THE BEHAVIORAL AND BRAIN SCIENCES (1978), 2, 215–260
Printed in the United States of America

The nature and plausibility of Cognitivism

John Haugeland
Department of Philosophy, University of Pittsburgh, Pittsburgh, Pennsylvania 15260

Abstract: Cognitivism in psychology and philosophy is roughly the position that intelligent behavior can (only) be explained by appeal to internal "cognitive processes," that is, rational thought in a very broad sense. Sections 1 to 5 attempt to explicate in detail the nature of the scientific enterprise that this intuition has inspired. That enterprise is distinctive in at least three ways: It relies on a style of explanation which is different from that of mathematical physics, in such a way that it is not basically concerned with quantitative equational laws; the states and processes with which it deals are "interpreted," in the sense that they are regarded as meaningful or representational; and it is not committed to reductionism, but is open to reduction in a form different from that encountered in other sciences. Spelling these points out makes it clear that the Cognitivist study of the mind can be rigorous and empirical, despite its unprecedented theoretical form. The philosophical explication has another advantage as well: It provides a much needed framework for articulating questions about whether the Cognitivist approach is right or wrong. The last three sections take that advantage of the account, and address several such questions, pro and con.

Keywords: explanation, cognition, information processing, reduction, methodology, computer models, philosophy of science

1. Systematic explanation

From time to time, the ills of psychology are laid to a misguided effort to emulate physics and chemistry. Whether the study of people is inherently "humanistic" and "soft" (Hudson, 1972), or whether states described in terms of their significance necessarily escape the net of physical law (Davidson, 1970), the implication is that psychology cannot live up to the standards of rigorous science, and perhaps cannot be a science at all. But science itself often leaves behind efforts to say what it can and cannot be. The Cognitive approach to psychology offers, I think, a science of a distinctive form, and thereby sidesteps many "philosophical" objections – including those born of a dazzled preoccupation with physics. In my first five sections I will try to characterize that form.

Science in general is an endeavor to understand what occurs in the world; hence explanation, which is essentially a means to understanding, has a pivotal importance. Scientific explanations differ from common sense explanations at least in being more explicit, more precise, more general, and more deliberately integrated with one another. Without attempting a full analysis, we can notice several broad characteristics which all scientific explanations share. They depend on specifying a range of features which are exhibited in, or definable for, a variety of concrete situations. They depend on knowing or hypothesizing certain regularities or relationships which always obtain in situations exhibiting the specified features. And they depend on our being able to see (understand), for particular cases, that since the specified features are deployed together in way X, the known regularities or relationships guarantee that Y. We then say that Y has been *explained* through an appeal to (or in terms of) the general regularities and the particular deployment of the features. The regularities and deployment appealed to have been presupposed by the explanation, and not themselves explained – though either might be explained, in turn, through appeal to further presuppositions.

Philosophers have coined the term *deductive-nomological* for explanations in which the presupposed regularities are formulated as laws (Greek: *nomos*), and for which the guarantee that Y will occur is formulated as a deductive argument from the laws plus statements describing the deployment X (Hempel and Oppenheim, 1949). It can be maintained that all scientific explanations are deductive-nomological, though in many cases that requires a counter-intuitive strain on the notion of "law." So to avoid confusion I will introduce some more restricted terminology, and at the same time illustrate several different ways in which the foregoing schematic remarks get fleshed out.

The most familiar scientific explanations come from classical mechanics. The situational features on which they depend include masses, inertial moments, distances, angles, durations, velocities, energies, and so on – all of which are quantitative, measurable parameters. The known regularities or relationships are expressed as equations (algebraic, vectorial, differential, etc.) relating the values of the various parameters in any given situation – e.g., $F = ma = dp/dt$. Usually some of the equations are designated laws and the others definitions, but there's a well-known trade-off in which are which. Equations are conveniently manipulable and combinable in ways that preserve equality; that is, other equations can be mathematically derived from them. The standard form of an explanation in mechanics is such a derivation, given specified deployments of masses, forces, and what have you (see Newton's derivations of Kepler's laws). It is the derived equational relationships which are explained (or sometimes the actual values of some of the parameters so related, determined by plugging in the known values of others).

I use *derivational-nomological* for this special case form of deductive-nomological explanation – where the distinction of the special case is that the presupposed regularities are expressed as equational relationships among quantitative parameters, and the deduction is a mathematical derivation of other such equations (and then, perhaps, computing some of the values). Besides mechanics, fields as diverse as optics, thermodynamics and macroeconomics commonly involve derivational-nomological explanations.

But what is important here is that there are other forms or styles of explanation, even in advanced sciences. I will delineate

 0140–525X/78/JHAUG006$04.00/0

two such distinct styles, though I will not claim that the distinctions are sharp. The claim is rather that interesting differences can be characterized among prime examples, despite the fact that intermediate cases blur the boundaries. Only one of these further styles is relevant to cognitive psychology; I delineate them both because they are superficially similar, and easily confused. Thus explicitly distinguishing them permits a closer focus on the one we want. These distinctions are independent of anything peculiar to psychology, and I will draw them that way first, to keep separate issues as clear as possible.

Imagine explaining to someone how a fiber optics bundle can take any image which is projected on one end and transmit it to the other end. I think most people would come to understand the phenomenon, given the following points. (If I am right, then readers unfamiliar with fiber optics should still be able to follow the example): (i) the bundles are composed of many long thin fibers, which are closely packed side by side, and arranged in such a way that each one remains in the same position relative to the others along the whole length of the bundle; (ii) each fiber is a leak-proof conduit for light – that is, whatever light goes in one end of a fiber comes out the other end of the same fiber; (iii) a projected image can be regarded as an array of closely packed dots of light, differing in brightness and color; and (iv) since each end of each fiber is like a dot, projecting an image on one end of the bundle will make the other end light up with dots of the same brightness and color in the same relative positions – thus preserving the image.

Clearly that was not a derivational-nomological explanation. One could, with effort, recast it as a logical deduction, but I think it would lose more perspicacity than it would gain (diagrams would help much more). If we do not try to force it into a preconceived mold of scientific explanations, then several distinctive aspects stand out as noteworthy. First, what is explained is a disposition or ability of a kind of object (compare Cummins, 1975). Second, the explanation makes appeals (presuppositions) of two basic sorts: that the kind of object in question has a certain form or structure (compare Putnam, 1975), and that whatever is formed or structured in that way has certain dispositions or abilities. (The object is a bundle of "parallel" fibers, and each fiber is able to conduct light without leaking.) Third, any object structured in the presupposed way, out of things with the presupposed abilities, would have the overall ability being explained. That is, it doesn't matter how or why the fibers are arranged as they are, nor how or why they conduct light; these are simply presupposed, and they are sufficient to explain the ability to transmit images.

I call explanations of this style *morphological*, where the distinguishing marks of the style are that an ability is explained through appeal to a specified structure and to specified abilities of whatever is so structured. (These specifications implicitly determine the "kind" of object to which the explanation applies). In science, morphological explanations are often called "models" (which in this sense amount to specifications of structure), but that term is both too broad and too narrow for our purposes: Logicians have a related but different use for it, whereas few would call the fiber optics account a model.

On the other hand, the account of how DNA can replicate itself is called a model – the double-helix model – and it is morphological. Simplistically put, the structure is two adjacent strands of sites, with each site uniquely mated to a complementary one in the other strand. And the sites have the ability to split up with their mates and latch onto an exactly similar new one, selected from a supply which happens to be floating around loose. This process starts at one end of the double strand, and, by the time it reaches the other end, there are two double strands, each an exact replica of the original. At the opposite extreme of sophistication, an explanation of how cups are able to hold coffee would be morphological. The specified structure would be little more than shape, and the specified abilities of what is so structured would amount to rigidity, insolubility, and the like.

Now consider a case which is subtly but importantly different: an explanation of how an automobile engine works. As with morphological explanations, this one appeals to a specified structure, and to specified abilities or dispositions of what is so structured. But in addition, and so important as to dominate the account, it requires specification of a complexly organized pattern of interdependent interactions. The various parts of an engine do many different things, so to speak "working together" or "cooperating" in an organized way, to produce an effect quite unlike what any of them could do alone.

I reserve the term *systematic* for explanations of this style, where the distinction from morphological explanation is the additional element of organized cooperative interaction. Strictly, it is again an ability or disposition which gets explained, but the ordinary expression "how it works" often gives a richer feel for what's involved. A consequence of this definition is that objects with abilities that get systematically explained must be composed of distinct parts, because specifying interactions is crucial to the explanation, and interactions require distinct interactors. Let a *system* be any object with an ability that is explained systematically, and *functional components* be the distinct parts whose interactions are cited in the explanation. In a system, the specified structure is essentially the arrangement of functional components such that they will interact as specified; and the specified abilities of the components are almost entirely the abilities to so interact, in the environment created by their neighboring components. Note that what counts as a system, and as its functional components, is relative to what explanation is being offered. Other examples of systems (relative to the obvious explanations) are radios, common mousetraps, and (disregarding some messiness) many portions of complex organisms.

Fiber optics bundles and DNA molecules are deceptively similar to systems, because they have clearly distinct components, each of which contributes to the overall ability by performing its own little assigned "job." But the jobs are not interdependent; it is not through cooperative interaction that the image transmission or replication is achieved, but only an orderly summation of the two cents' worth from each separate fiber or site. In an engine, the carburetor, distributor, spark plugs, and so forth, do not each deliver a portion of the engine's turning, in the way that each site or fiber contributes a portion of the replication or image. The job metaphor can be expanded to further illustrate the difference. In old fashioned plantation harvesting, each laborer picked a portion of the crop (say one row), and when each was done, it was all done. But at a bureaucratic corporation like General Motors, comparatively few workers actually assemble automobiles; the others make parts, maintain the factories, come up with new designs, write paychecks, and so on. All of these tasks are prerequisite to continued production, but only indirectly, through a complex pattern of interdependencies. A system is like a bureaucratic corporation, with components playing many different roles, most contributing to the final outcome only indirectly, via the organized interactions.

I have described three different styles of explanation, each of which can be scientifically rigorous and respectable. They are all abstract or formal, in that they all "abstract" certain features and regularities from a variety of concrete situations, and then show how the resulting "forms" make certain properties or events intelligible in all such situations. But they differ notably in the nature of the abstract forms they specify, at least in clear cases. Only the derivational-nomological style puts an explicit emphasis on equations of the sort that we usually associate with scientific laws. But I shall claim that only the systematic style is directly relevant to cognitive psychology. Thus the charge of slavishly imitating mathematical physics does not apply to Cognitivism, and it doesn't matter that quantitative equational "laws of behavior" seem to be few and far between. Many of the points I have made have been made before (Cummins, 1975; Putnam, 1975, ch. 14; Dennett, 1971; Simon, 1969; Fodor, 1965), but no one, to my knowledge, has previously distinguished mor-

phological and systematic explanation. The importance of that distinction will emerge in section 4.

2. Systematic reduction

Traditional philosophical concerns for the unity of science and for the metaphysical doctrine of materialism (i.e., that everything is "ultimately just" matter in motion) customarily lead to questions about scientific reduction. Psychological concepts and theories are prime targets for such questions because they are not, at first glance, materialistic. This is not the place for a full discussion of the problem of reduction, but my position about the nature of Cognitivism will have several specific implications which should be pointed out. Some of these derive from the suggestion that Cognitivist explanation is systematic, and those can be considered independently of issues peculiar to psychology.

An aspect common to all the explanations discussed in the last section (indeed, to all explanations) is that they presuppose some things in the course of explaining others. More particularly, they presuppose certain specified general regularities, which are appealed to, but not themselves explained. But such regularities often can be explained, by appeal to others which are more "basic." Such further explanation is *reduction*, though obviously it counts as reduction only relative to the explanations whose presupposed regularities are being explained. This is a fairly broad definition of reduction, and includes cases which aren't very exciting in form. Thus, Newton's derivation of Kepler's laws counts as a reduction of Kepler's explanations of planetary positions.

A more famous reduction in classical physics, and one with a more interesting form, was that of thermodynamics to statistical mechanics. In outline, the values of the parameters occurring in the equations of thermodynamic theory were found (or hypothesized) to correlate with quantities definable statistically in terms of the mechanical parameters of groups of atoms. For example, the absolute temperature of a region was found to be proportional to the average kinetic energy of the atoms in that region. Such correlations are expressed in specific equations called "bridge equations." It then turned out that the laws of thermodynamics could be mathematically derived from the laws of mechanics, some plausible statistical assumptions, and these bridge equations. The effect was to explain the regularities which were presupposed by thermodynamic explanations – in other words, to reduce thermodynamics.

Reductive explanations which explain the equational laws presupposed by derivational-nomological explanations I call *nomological reductions*. Note that the definition refers to the style of explanation being reduced, not to the style of the reducing explanation. The reduction of thermodynamics is often cited as a paradigm of scientific reduction, as if all others should have a similar structure. But a moment's reflection shows that this structure only makes sense if the explanation being reduced is derivational-nomological; otherwise there would be no equational laws to derive, and probably no quantitative parameters to figure in bridge equations.

The regularities presupposed by morphological and systematic explanations are mainly the specified dispositions or abilities of whatever is structured in the specified way. Hence, *morphological* and *systematic reductions* (which are pretty similar) are explanations of those abilities. Such reducing explanations can themselves be of various styles. Thus, an explanation of how thin glass fibers can be light conduits would be, I think, borderline between morphological and derivational-nomological. But the explanation of how DNA sites can do the things appealed to in the replication explanation is very complex; and, for all I know, it is systematic.

In explaining a system, almost all of the abilities presupposed are abilities of individual components to interact with certain neighboring components in specified ways. Since intricate, interdependent organization is the hallmark of systems, the abilities demanded of individual components are often enough themselves rather sophisticated and specialized. Conversely, since systems typically have abilities strikingly different from those of any of their separate components, systematic organization is a common source of sophisticated and specialized abilities. These considerations together suggest that very elaborate systems could be expected to have smaller systems as functional components. And frequently they do – sometimes with numerous *levels* of systems within systems. For example, a car's ignition system is a component in the engine system, and the distributor system is a component in that. Such a multilevel structure of nested systems is a *systematic hierarchy*. (See Simon, 1969, for further discussion of hierarchical organization.)

So a systematic reduction of the highest system in a systematic hierarchy would involve systematic explanations of the specified interactive abilities of its functional components; and perhaps likewise for reductions of those, and so on. Only at the lowest level would systematic reductions be a different style of explanation (typically morphological; compare the explanation of a crankshaft or piston to that of a coffee cup). Since any scientific reduction is also a scientific explanation, it will explicitly presuppose certain regularities, which can be enquired after in turn. At any given time, however, some regularities will not be explainable. Modern wisdom has it that in the golden age these will include only the "fundamental" laws of physics, all others being reducible to them (perhaps through many stages of reduction). A sequence of reductions taking the presuppositions of an explanation all the way to physics is a *complete reduction*. A complete reduction of psychology is one of the dreams of unified science.

A common misconception is that reductions supplant the explanations they reduce – that is, render them superfluous. This is not so. Consider the fiber optics reduction. There could be any number of different explanations for why different kinds of fibers can conduct light; thus glass threads, with variable index of refraction versus radius, would call for a different explanation than hollow silver tubes. But those are irrelevant to the explanation of how the bundle transmits images. The latter takes light conduction in the fibers for granted and goes on to tell us something new. This something new would be lost if we settled exclusively for explanations of light conductivity; and on the other hand, it would not be lost (given the original morphological explanation) even if light conductivity were totally inexplicable. The two explanations are quite independent, even though one is of the presuppositions of the other (compare Putnam, 1975, Ch. 14).

The main point of this section has been that reductions, like explanations, are not all alike. Hence, the reduction of thermodynamics cannot serve as a universal paradigm, despite its ubiquitous use as an example. In particular, if I am right that Cognitivist explanation is systematic, then any reduction of Cognitivism would be systematic reduction (a point to be taken up further in section 5). This means at least that Cognitivists are not interested in "psycho-physical bridge" equations (*pace* Fodor, 1974), nor are they worried if none are possible (*pace* Davidson, 1970).

3. Intentional interpretation

Because the study of the mind presents special scientific difficulties all of its own, I have so far mentioned psychology only incidentally. At the heart of these special difficulties is the problem of "significance" or "meaningfulness." Large portions of human behavior, preeminently linguistic behavior, are meaningful on the face of it, and a larger portion still is "rational" or "intelligent" in a way that involves significance at least indirectly. Yet meaningfulness is a slippery notion to pin down empirically, and there are conceptual difficulties in connecting "meanings" with the physical order of cause and effect. So serious are the

problems that some investigators have even tried to study behavior entirely without regard to its significance, but their achievements have been narrow and limited. Cognitivism, on the other hand, gives the meanings of various states and processes a central importance. In this section, I will show how that can be compatible with the rigorous demands of empirical science.

I take my cue from the pioneering work of Quine, and the refinements it has inspired (Quine, 1960; Davidson, 1970, 1973; Harman, 1973; compare also Sellars, 1963, Ch. 11; Dennett, 1971). Quine's original concern was the translation of utterances in totally alien languages; since Cognitivism's topic is broader, we generalize "translation" to "intentional interpretation" and "utterance" to "quasilinguistic representation." These now must be explicated.

Suppose we come upon an unfamiliar object – a "black box" – which someone tells us plays chess. What evidence would it take to convince us that the claim was empirically justified? It is neither necessary nor sufficient that it produce tokens of symbols in some standard chess notation (let alone, physically move the pieces of a chess set). It is not sufficient because the object might produce standard symbols, but only in a random order. And it is not necessary, because the object might play brilliant chess, but represent moves in some oddball notation.

So it is up to the person who claims that it plays chess to tell us how it represents moves. More particularly, we must know what in its behavior to count as its making a move, and how to tell what move that is. Further, we must know what effects on it count as opponents' moves, and how to tell what moves they count as. Succinctly: we must know what its inputs and outputs are, and how to interpret them. Note that the inputs and outputs must be of some antecedently recognizable or identifiable types, and the interpretations of them must be according to some antecedently specifiable regular scheme; otherwise, we will suspect that the "interpretation" is being made up along the way, so as to make things come out right.

Of course, simply specifying the interpretation does not convince us that the object really plays chess. For that we would need to watch it play a few games – perhaps with several opponents, so we're sure there's no trick. What will count as success in this test? First, each output that the object produces must turn out, under the specified interpretation, to be a legal move for the board position as it stands at that time. Second, depending on how strictly we distinguish blundering from playing, the moves must be to some extent plausible (the hypothesis is only that it plays, not that it plays well). If the object passes this test in a sufficient variety of cases, we will be empirically convinced that it is indeed a chess player.

Further, when the object passes the test, the original interpretation scheme is shown to be not merely gratuitous. This is important because, in themselves, interpretation schemes are a dime a dozen. With a little ingenuity, one can stipulate all kinds of bizarre "meanings" for the behavior of all kinds of objects; and insofar as they are just stipulations, there can be no empirical argument about whether one is any better than another. How would you test, for example, the claims that producing marks shaped like "Q-B2" represented ("meant"): (i) one (or another) particular chess move, (ii) the solution of a logic problem, or (iii) a scurrilous remark about Queen Elizabeth and Bishop Sheen? Nothing observable about those marks in themselves favors one rendition over another. But one can further observe when and where the marks are produced, in relation to others produced by the same object, and in relation to the object's inputs. If those relationships form a pattern, such that under one interpretation the observed outputs consistently "make reasonable sense" in the context of the other observed inputs and outputs, while under another interpretation they don't, then the first interpretation scheme as a whole is observably "better" (more convincing) than the second. In our example, the pattern amounts to playing legal and plausible chess games, time after time. None (or at most very few) of the countless other conceivable interpretations of the same marks would make such sense of the observed pattern, so the given interpretation is empirically preferable.

The problem now is to generalize the points made about this specific example. I believe there are principled limits to how precisely such a generalization can be stated; but let us proceed with a few definitions, relying on intuitions and examples to keep them clear.

1. A set of types is *uniquely determinable* relative to a specified range of phenomena iff:

   i. for almost every phenomenon in that range one can unequivocally determine whether it is a token (instance) of one of the types, and if so, which one; and

   ii. no phenomenon is ever a token of more than one type.

(Compare Goodman, 1968, Ch. 4; Quine, 1960, section 18.)

2. An *articulated typology* (relative to a range of phenomena) is an ordered pair of uniquely determinable sets of types such that:

   i. tokens of types in the second set (= *complete types*) are composed of one or more tokens of types in the first set (= *simple types*); and

   ii. no token of a simple type ever actually occurs in the specified range of phenomena except as a component of a complete type.

For example, suppose a sheet of paper has a chess game recorded on it in standard notation (and has no other markings but doodles). Then relative to the marks on that page, the alphameric characters of chess notation are the simple types of an articulated typology, and the sequences of characters that would canonically represent moves (plus odds and ends) are the complete types. Note that definitions of complete types may include specifications of the order in which they are composed of simple types, and that in general this order need not be merely serial.

3. An *intentional interpretation* of an articulated typology is:

   i. a regular general scheme for determining what any token of a complete type means or represents, such that:

   ii. the determination is made entirely in terms of:

      a. how it is composed of tokens of simple types; and

      b. some stipulations about ("definitions" of) the simple types.

("Intentional" is a philisopher's term for "meaningful" or "representational.")

4. A *quasilinguistic representation* is a token of a complete type from an intentionally interpreted articulated typology.

(Compare "structured description," Pylyshyn, in press.) Obviously, the identity of a quasilinguistic representation is relative to the specified typology and interpretation, and hence also to a specified range of phenomena. The complete types in the chess notation typology are quasilinguistic representations (of moves), relative to the chess interpretation.

I am unable to define either "mean" or "represent," nor say in general what kinds of stipulations about simple types (3-ii-b, earlier) are appropriate. In practice, however, it is not hard to give clear intentional interpretations; there are two common ways of doing it. The first is translation into some language or notation that we, the interpreters, already understand. Thus a manual might be provided for translating some strange chess notation into the standard one. The second is giving an "intended interpretation," in roughly the logicians' sense. Thus, a function can be defined from a subset of the simple types onto some domain – say, chess pieces and board squares; then the meanings of tokens of complete types (e.g., what moves they represent) are specified recursively in terms of this function, plus the roles of other simple types (such as punctuation) characterized implicitly by the recursion.

Definitions 1 to 4 were all preparatory for the following.

5. An object is interpreted as an *intentional black box* (an IBB) just in case:

   i. an intentionally interpreted articulated typology is specified relative to the causal influences of its environment on it – resulting quasilinguistic representations being *inputs;*

   ii. likewise for *outputs,* relative to its causal influences on its environment; and

   iii. it is shown empirically that under the interpretations the actual outputs consistently make reasonable sense in the context (pattern) of

actual prior inputs and other actual outputs.

(It will sometimes be convenient to use the term "IBB" on the assumption that such an interpretation can be given, even though the specifics are not known.) The chess player with which this section began is an IBB; so are adding machines, logic problem solvers, automated disease diagnosers, and (applying the definitions flexibly) normal people.

There are three problems with this definition that need immediate comment. First, "making reasonable sense" under an interpretation is not defined – and I doubt that it can be. Again, however, it is seldom hard to recognize in practice. Often, explicit conditions can be stated for making sense about certain problem domains or subject matters; these I call *cogency conditions*. For the chess player, the cogency condition was outputting legal and plausible moves in the context created by the previous moves. For interpreting an object as an adding machine, the condition is giving correct sums of the inputs; for a disease diagnostician it is giving good diagnoses relative to the symptoms provided. Various authors have tried to give completely general cogency conditions for interpreting creatures as language users (Wilson, 1959; Quine, 1960, Ch. 2; Lewis, 1974; Grandy, 1973; Davidson, 1973). For reasons beyond the scope of this discussion, I don't think any of these succeed. But it doesn't matter much in actual field or laboratory work, because by and large everyone can agree on what does and doesn't make sense.

Second, if one is knee-jerk liberal about what makes reasonable sense, then all kinds of objects can be trivially interpreted as IBBs. Thus, a flipped coin might be interpreted as a yes-no decision maker for complex issues tapped on it in Morse code. I will assume that such cases can be ignored.

Third, and most serious, the requirement that inputs and outputs be quasilinguistic representations appears to rule out many perceptions and actions. In at least some cases, this problem can be handled indirectly. Suppose an alleged chess player used no notation at all, but had a TV camera aimed at the board and a mechanical arm which physically moved the pieces. The problem of showing that this device indeed plays chess is essentially the same as before: It must consistently make legal and plausible moves. This succeeds, I think, because we can give quasilinguistic descriptions of what it "looks" at and what it does, such that if they were the inputs and outputs the object would count as an IBB. In such cases, we can enlarge our interpretation, and say that the object perceives and acts "under those descriptions" (sees that . . . , intends that . . . , etc.), and regard the descriptions as inputs and outputs. Where this strategy won't work, my definition won't apply.

In this section I have addressed the question of how meaningfulness or significance can be dealt with empirically. In brief, the idea is that although meaningfulness is not an intrinsic property of behavior that can be observed or measured, it is a characteristic that can be attributed in an empirically justified interpretation, if the behavior is part of an overall pattern that "makes sense" (e.g., by satisfying specified cogency conditions). In effect, the relationships among the inputs and outputs are the only relevant observational data; their intrinsic properties are entirely beside the point, so long as the relationships obtain. But the fact that they have some characteristics or other, independent of the interpretation (that is, they are causal interactions with the environment), means that there is no mystery about how states with significance "connect" with the rest of nature (Davidson, 1970). The upshot is that a psychological theory need not in principle ignore meaningfulness in order to maintain its credentials as empirical and scientific.

4. Information processing systems

The last section showed only that there is an empirically legitimate way to talk about significances in scientific theories. It did not say anything about what kind of scientific account might deal with phenomena in terms of their "meanings." To put it another way, we only saw how the notion of IBB could make good empirical sense, not how anything could be explained. Yet an IBB always manages to produce reasonable outputs given its inputs; and that's a fairly remarkable ability, which cries out for explanation. There may be many ways to explain such an ability, but two in particular are relevant to Cognitivism. One will be the subject of this section, and the other of the next.

If one can systematically explain how an IBB works, without "de-interpreting" it, it is an *information processing system* (an IPS). By "without de-interpreting," I mean explaining its input/output ability in terms of how it would be characterized under the intentional interpretation, regardless of whatever other descriptions might be available for the same input and output behavior. For example, if our chess player is an IPS, that means there is a systematic explanation of how it manages to come up with legal and plausible moves as such, regardless of how it manages to press certain type bars against paper, light certain lights, or do whatever it does that gets interpreted as those moves.

In a systematic explanation, the ability in question is understood as resulting from the organized, cooperative interactions of various distinct functional components, plus their separate abilities. Further, whatever result it is that the object is able to yield (in this case the IBB outputs), is typically delivered directly by some one or few of the functional components. Now, since we're not de-interpreting, those few components which directly deliver the outputs of the IPS must have among their presupposed abilities the ability to produce the outputs as interpreted. But if attributing this ability to those components is to make good empirical sense, then they must be IBBs themselves. Hence the effects on them by their functional neighbors in the system (the interactions appealed to in the explanation) must be their IBB inputs, which means that they too are dealt with as interpreted. But since these inputs are at the same time the effects delivered by other components, those other components must be able to deliver effects (outputs) under an interpretation. Consequently, they also – and by the same argument, all the functional components of an IPS – must be IBBs.

Moreover, all the interpretations of the component IBBs must be, in a sense, the same as that of the overall IBB (= the IPS). The sense is that they must all pertain to the same subject matter or problem. This actually follows from the preceding argument, but an example will make it obvious. Assuming that the chess playing IBB is an IPS, we would expect its component IBBs to generate possible moves, evaluate board positions, decide which lines of play to investigate further, or some such. These not only all have to do with chess, but in any given case they all have to do with the same partially finished game of chess. By contrast, components interpreted as generating football plays, evaluating jockeys, or deciding to pull trump could have no part in explaining how a chess player works.

Still, the sense in which the interpretations have to be the same is limited. First, of course, the types which get interpreted can vary throughout; they might be keyboard characters in one case, electric pulses in another, and so forth. More important, the internal "discourse" among component IBBs can be in a richer "vocabulary" than that used in the overall inputs and outputs. Thus, chess player inputs and outputs include little more than announcements of actual moves, but the components might be engaged in setting goals, weighing options, deciding which pieces are especially valuable, and so on. Even so, they all still pertain to the chess game, which is the important point. (The importance will become clearer in section 5).

It is natural in a certain way to seek a systematic explanation of an IBB's input/output ability. Seeing this is to appreciate one of the essential motivations of Cognitivism. The relevant ability of an IBB is to produce reasonable outputs relative to whatever inputs it happens to get from within a wide range of possibilities. In a broad sense of the term, we can think of the actual inputs as posing "problems," which the IBB is then able to solve. Now only certain outputs would count as reasonable solutions to any

given problem, and those are the ones for which some kind of reasonable argument or rationale can be given. (Cogency conditions are typically spelled out as a relevant rationale for certain outputs as opposed to others, given the inputs). An argument or rationale for a solution to a problem amounts to a decomposition of the problem into easier subproblems, plus an account of how all the subsolutions combine to yield a solution of the overall problem. (How "easy" the subproblems have to be is, of course, relative to the context in which the rationale is required). The point is that the separate IBB components of the IPS can be regarded as solving the easier subproblems, and their interactions as providing the combination necessary for coming up with the overall solution. The interactions in general must be organized and "cooperative" (i.e., systematic) because rationale considerations and relationships generally "combine" in complexly interdependent and interlocking ways. (This is why the systematic/morphological distinction is important.)

So, the interacting components of an IPS "work out," in effect, an explicit rationale for whatever output they collectively produce. And that's the explanation for how they manage to come up with reasonable outputs; they, so to speak, "reason it through." This also is the fundamental idea of cognitive psychology: intelligent behavior is to be explained by appeal to internal "cognitive processes" – meaning, essentially, processes interpretable as working out a rationale. Cognitivism, then, can be summed up in a slogan: the mind is to be understood as an IPS.[1]

This suggestion stands on two innovative cornerstones, compared to older notions about what psychology should look like as a science. The first is that psychological explanation should be systematic, not derivational-nomological; hence, that psychology is not primarily interested in quantitative, equational laws, and that psychological theories will not look much like those in physics. The second is that intentional interpretation gives an empirically legitimate (testable) way of talking and theorizing about phenomena regarded as meaningful; hence, that psychology does not have to choose between the supposedly disreputable method of introspection, and a crippling confinement to purely behavioral description. Together they add up to an exciting and promising new approach to the study of the mind.

5. Intentional reduction

The abilities of component IBBs are merely presupposed by an IPS explanation. That explanation can be systematically reduced – in the sense of section 2, by turning one's attention to explaining those component abilities. If it happens that the components are themselves IPSs, then reduction can proceed a step by appealing to the organized interactions and abilities of still smaller component IBBs, and so on. An extension of the argument in the last section shows that all the IBB components at all the levels in such a hierarchy must be interpreted as having the same subject matter; for example, all their inputs and outputs pertain to the same game of chess, or whatever.

Obviously, then, a complete reduction down to physics (or electronics or physiology) would have to involve some further kind of step; that is, eventually the abilities of component IBBs would have to be explained in some other way than as IPSs. By definition, IPS explanation does not involve de-interpretation. Explanation of an IBB's input/output ability that does involve de-interpretation I call explanation by *instantiation*. We shall see that instantiation has two importantly distinct forms.

An object of the sort computer engineers call an "and-gate" is a simple IBB. It has two or more input wires, and a complete input type is (for example) a distribution of positive and negative voltages among these wires. It has one output wire, and is constructed electronically to put a positive voltage in this wire if and only if all the input voltages are positive; otherwise it puts out a negative voltage. Now the cogency condition for a proposition conjoiner is that it give the truth-value "true" if and only if all the conjoined propositions are true; otherwise it gives "false." Since this truth function for "and" is isomorphic to the electrical behavior of the object (taking positive voltage as "true," and negative as "false"), the object can be interpreted as an and-gate.

But to explain how the object manages to satisfy the prescribed cogency conditions, one would not look for component IBBs interpretable as "reasoning the problem through." Rather, one would de-interpret and explain the electrical behavior in terms of the electric circuitry and components. The electrical circuit might well be a system, but it would not be an IPS. Since the first step of the explanation is de-interpretation, it is an explanation by instantiation; I call it *physical instantiation* because the remainder of it is expressed in physical terms.

Not all instantiations, however, are physical instantiations. For example, computer-based chess players are generally written in a programming language called LISP, in which the inputs and outputs of subroutines are interpreted as constructing and manipulating complex lists. So interpreted, these subroutines are IBBs, but their subject matter is not chess. What happens, however, is that the input/output constraints (cogency conditions) on the lowest level components in the chess related hierarchy are isomorphic to the constraints on IBBs built up in LISP.[2] Thus, the required abilities of bottom-level chessplayer components can be explained by de-interpreting (or re-interpreting) them as IBBs solving problems about list-structures – IBBs which can then be understood as IPS's working through the rationale for the LISP problem. This, too, is reduction by instantiation, but I call it *intentional instantiation*, because the redescribed ability is still an IBB ability, just about a different subject matter.

Actually, in a complete reduction of a fancy computer program, there can be several stages of intentional instantiation. Thus, LISP languages are generally written (compiled) in still more basic languages – say, ones in which the only IBB abilities are number-crunching and inequality testing (the conditional branch). The last intentional instantiation is in a primitive "machine language," so-called because that is the one which is finally reduced by physical instantiation. The real genius of computer science has been to design ever more sophisticated languages which can be compiled or intentionally instantiated in cruder existing languages. If it weren't for intentional instantiation, machines built of flip-flops and the like would hardly be candidates for artificial intelligence.

It is easy to confuse the maneuver of explaining an IBB by intentional instantiation with that of explaining it as an IPS. The essential difference is the re-interpretation – or, intuitively, the change in subject matter. Since I have already used "change of level" to describe the move from IPS to its separate components, I will use "change of *dimension*" to describe the move of de-/re-interpretation involved in an instantiation. One can think of the many dimensions in a sophisticated "system" as forming a hierarchy, but dimension hierarchies should not be confused with the earlier level hierarchies. There can be different level hierarchies on different dimensions, but they are "orthogonal" rather than sequential. That is, it's a mistake to think of the lowest level on one dimension as a higher level than the highest level on a lower dimension. Thus, an and-gate is not a higher level component than a disk memory; they are components on different dimensions, and hence incomparable as to level.[3]

The importance of keeping dimensions and levels straight is illustrated by the following nonsequitur of Lucas' (Lucas, 1964). Recast in our terminology, Lucas argued that Gödel's incompleteness result implied that the mind of a good mathematician could not be understood as an IPS. The reasoning was that any IPS could be simulated by a Turing machine of the sort that can also be interpreted as proving theorems in a formal system. But Gödel proved that any formal system rich enough to include arithmetic would be able to express truths which were not provable in the system (Gödel, 1931; Nagel and Newman, 1958).

Moreover, for any particular formal system, a good mathematician would be able knowingly to construct such a truth – which the corresponding theorem-proving Turing machine in principle could not do. Hence, for any given Turing machine, a good mathematician can do something which the former cannot, and therefore they cannot be the same. The error here is in neglecting that an interpretation of an object as a mathematician and an interpretation of it as a theorem-proving Turing machine would be on vastly different dimensions. The outputs put forward as proofs and truth claims on one bear no interesting relation to those put forward on the other.

In this section, I have outlined what a reduction of cognitive psychology to the relevant physical dimension theory would look like. I have not argued that Cognitivism *per se* is committed to such reducibility. It would be theoretically consistent to maintain that at some bottom level the presupposed IBB abilities were simply not explainable (much as physics cannot explain its fundamental laws). Nevertheless, I suspect that many investigators would strongly resist such a suggestion, and would feel their work was not done until the reduction was complete.

6. Fallacious supporting arguments

In sections 1 through 5, I have given a general characterization of the Cognitivist approach to psychology, and its possible reduction. In so doing, I have shown how it is innovatively different from earlier approaches more captivated by the image of physics, and how it can be unimpeachably rigorous and empirical all the same. However, it seems to me that the eventual success of this program, for all its attractiveness, is still very much in doubt. In the remaining three sections, I hope to make clear my reasons for caution – taking as much advantage as possible from the explicit characterization just completed. I will begin in this section by pointing out the flaws in two seductive general arguments to the effect that some Cognitivist theory or other must be right.

The first argument is directed more specifically at the systematicity cornerstone, though as we have seen, the two cornerstone innovations go hand in hand (see the end of section 4). It goes like this: we know that the nervous system is composed of numerous distinct and highly organized "functional components" – namely neurons; and (assuming materialism) there is every reason to believe that the human IBB is somehow instantiated in the nervous system. So, all that remains to be found are how the neurons are grouped into higher level components, how the first instantiation proceeds, how the lowest components on that dimension are grouped into higher components, what the next instantiation is, and so on. That is, we need only "build back up" the intentional and systematic reductions described in sections 2 and 5, until we reach the overall IBB. That's an enormous task, of course, but since we know there are organized components at the bottom, we know in principle it can be done.

Formally this argument is circular; the reductions mentioned in describing the "building back up" presuppose the very systematicity that the argument is supposed to prove. But the idea behind the reasoning is so attractive that it is tempting to think that the circularity is an artifact of the formulation, and that a better version could be found. To see that this is not so, we must expose in detail the real basis of the formal circularity.

As we observed in section 1, scientific explanation is essentially a route to understanding; and the understanding is achieved in part through specifying certain features and regularities that are common to the range of situations where that kind of explanation applies. The demands of rigor and explicitness that distinguish some explanations as scientific require that the features and regularities specified "encompass" or "encapsulate" every consideration that is relevant to understanding the phenomenon being explained. In a way, the explanatory insight derives precisely from the realization that these few specific features and regularities are all you need to know, in order to be sure that phenomenon *Y* will occur; everything else is extraneous. Thus, the beauty of Newton's mechanics is that a few parameters and equational laws encapsulate everything that is relevant to the motions of a great many bodies. For example, the colors, textures, personalities, and so on of the planets can all safely be ignored in predicting and understanding their positions as a function of time.

In a systematic explanation, a comparable encapsulation is achieved in the specification of a few determinate modes of interaction among a few distinct components with particular specified abilities. Indeed, finding interfaces among portions of an object, such that this kind of encapsulation is possible, is the fundamental principle of individuation of functional components – and hence a *sine qua non* of systematic explanation. For example, dividing the interior of a radio (or engine) into adjacent one-millimeter cubes would not be a decomposition into functional components; and the reason is exactly that the resulting "interfaces" would not yield any dividend of encapsulating what's relevant into a few highly specific interactions and abilities. By contrast, a resistor can be a functional component, because nothing about it matters except the way it resists the flow of electricity from one of its leads to the other. (Of course, resistors ought not to melt, multiply, or eat diodes, but such injunctions are relevant not so much to how the radio works as to how the various components maintain their required abilities; i.e., they are a concern of reduction.)

So if neurons are to be functional components in a physiological system, then some specific few of their countless physical, chemical, and biological interactions must encapsulate all that is relevant to understanding whatever ability of that system is being explained. This is not at all guaranteed by the fact that cell membranes provide an anatomically conspicuous gerrymandering of the brain. More important, however, even if neurons were components in some system, that still would not guarantee the possibility of "building back up." Not every contiguous collection of components constitutes a single component in a higher level system; consolidation into a single higher component requires a further encapsulation of what's relevant into a few specific abilities and interactions – usually different in kind from those of any of the smaller components. Thus the tuner, pre-amp and power amp of a radio have very narrowly specified abilities and interactions, compared to those of some arbitrary connected collection of resistors, capacitors, and transistors. The bare existence of functionally organized neurons would not guarantee that such higher level consolidations were possible. Moreover, this failure of a guarantee would occur again and again at every level on every dimension. There is no way to know whether these explanatory consolidations from below are possible, without already knowing whether the corresponding systematic explanations and reductions from above are possible – which is the original circularity.

The second argument I will refute starts from the top rather than the bottom, and is directed primarily at the intentional interpretation cornerstone, with its associated idea of "working out the rationale." Formally this argument amounts to the challenge: What else could it be? – but it is much more persuasive than that brazen rendition suggests. If one disregarded the intentional interpretation of any sophisticated IBB, it would be quite incredible to suggest that there was some elegant relation between the particular set of influences from the environment that we call inputs, and the particular set of influences on the environment that we call outputs. The relevant actual pattern can hardly even be described except in some way that is tantamount to specifying the cogency conditions which the object in fact meets. But since what we observe is that the object consistently meets these otherwise quite peculiar conditions, and since the conditions themselves are typically made explicit by spelling out some rationale, what else could explain the observations than that the object works the rationale out? How else would it happen to come upon those particular outputs time after time?

To show that a "what else could it be?" argument is inconclusive one need only come up with a conceivably viable alternative; one need not make a case that the alternative is in fact more probable, just that it's viable. I will try to construct such an alternative, drawing on recent neurophysiological speculations about holographic arrangements and processes (van Heerden, 1968; Pribram, 1971, 1974; Pribram, et al., 1974; Pollen and Taylor, 1974). Fairly detailed hypothetical models have been proposed for how holograms might be realized in neuronal structures (Kabrisky, 1966; Baron, 1970; Cavanagh, 1972); and there is some empirical evidence that some neurons behave in ways that would fit the models (Campbell, 1974; Pollen and Taylor, 1974; and compare Erickson, 1974).

Optical holograms are photographs of interference patterns, which look kind of like the surface of a pond that has just had a lot of pebbles thrown in it. But they have some interesting properties (Leith and Upatnieks, 1965; Herriott, 1968; Cathey, 1974). First, they are prepared from the light bouncing off an ordinary object, and can subsequently be used to reconstruct a full three-dimensional colored image of that object. Second, the whole image can be reconstructed from any large enough portion of the hologram. (That is, there's no saying which portion of the hologram "encodes" which portion of the image). Third, a number of objects can be separately recorded on the same hologram, and there's no saying which portion records which object. Fourth, if a hologram of an arbitrary scene is suitably illuminated with the light from a reference object, bright spots will appear indicating (virtually instantaneously) the presence and location of any occurrences of the reference object in the scene (and dimmer spots indicate "similar" objects). So some neurophysiological holographic encoding might account for a number of perplexing features of visual recall and recognition, including their speed, some of their invariances, and the fact that they are only slightly impaired by large lesions in relevant areas of the brain.

What matters to us is that a pattern-recognizer based on these principles would not (or need not) be an IPS. There are no distinct functional components whose relevant interactions are confined to intentionally interpreted articulated typologies. That is, there is nothing going on which can be regarded as "working out a rationale" with quasilinguistic representations. By contrast a typical computer-based pattern-recognizer is an IPS. Thus, searching for discontinuities in luminance gradients, proposing that they are edges, checking for connexity among proposed edges, hypothesizing invisible edges so as to complete coherent objects, and so on are all rational procedures relative to the "problem" of identifying objects (see, e.g., Minsky and Papert, 1972).

The neurophysiologists cited have rightly confined their speculations to recognition and recall processes, because there one at least has shreds of evidence to work with. (But see Yevick, 1975, for a mathematician's tentative proposal of a holographically based "logic.") We, however, who are answering a "what else could it be?" argument needn't be so circumspect.

Another interesting property of optical holograms is that if a hologram of two objects is illuminated with the light from one of them, an image of the other (absent) object appears (Gabor, 1969; Firth, 1972). Thus, such a hologram can be regarded as a kind of "associator" of (not ideas, but) visual patterns. So imagine a set of such associated patterns, in which the first member of each is a common important substructure in chess positions, and the other is one or two moves which are generally powerful or dangerous around such structures. It seems to me that a set-up like that could be a nearly instantaneous "plausible move generator" for chess positions in general. In fact, it would mesh nicely with much of what is known about how human chess players perceive the board and their options (de Groot, 1965; Hearst, 1967; Frey and Adesman, 1976). Moreover, I don't see why ramification and complication of similar arrangements couldn't yield an object which generated actual moves. (How, after all, do masters play "three-second" chess when there's practically "no time to think?") Implementation of such a device by known optical means would be difficult or impossible, but I think it fair to point out also some limitations of the analogy. Optical holograms are all two-dimensional and static, whereas the brain is three-dimensional and dynamic. Who knows what could happen among holograms in intersecting planes, or even decomposed into dots and scattered about in volumes – especially if they dynamically modified one another in nonchaotic ways? Not so many years ago, no one could have imagined, except with the vaguest handwaves, what two-dimensional static holograms can do. Perhaps living brain homograms could do much more.

Again, the plausible move generator (or even whole chess player) that I have speculatively outlined would not be an IPS. Nothing in it "reasons through" the various possible moves and countermoves which would constitute the rationale for the legality and plausibility of the resulting output. It does not even contain a representation of the rules of the game, which would be unthinkable in a chess-playing IPS. But a chess player is a paradigm example of the kind of case to which the "what else could it be?" argument was supposed to apply. (Indeed, it's no accident that this example was so expositorily convenient in sections 3 through 5). I therefore take that argument to be refuted. I am not envisioning, of course, that humans (chess players included) engage in no cognitive "reasoning a problem through"; introspection, for all its ills, is enough to scotch that. But cognitive psychology is exciting and important for the unobvious thesis that cognitive information processing can explain much more than deliberate cogitation and reasoning; and for that larger thesis, the argument considered is inconclusive.

This last observation should put the whole present section in perspective. All I claim is that a few commonplace assumptions will not suffice to demonstrate that Cognitivism is the right approach to psychology. That should offend no one, since it only means that the position is not trivial and obvious – as clearly it isn't.

7. Potentially serious hurdles

In this section, I want to mention three issues which it seems to me may be serious hurdles for Cognitivism – serious in the sense of being equally hard to duck or get over. They are: moods, skills, and understanding. I cannot prove that Cognitivist accounts of these phenomena are impossible. My aim is rather to show that such accounts are going to be required if Cognitivism is to succeed, and that it's dubious whether they will be possible.

I will try to illustrate the nature of the difficulty with moods by contrasting it with another, which is superficially similar but more plausibly duckable. There is a long and tortured tradition in philosophy for distinguishing two kinds of mental phenomena: roughly, cognitive or intellectual states vs. felt qualities or the purely sensuous given (see Sellars, 1963, ch. 5, for a discussion of this distinction in the context of a different issue). Paradigm "felt qualities" would be pains or mere awarenesses of present red (not categorized or conceptualized as such). Several recent articles have argued that such states have some kind of determinate immediate character which is independent of any interpretation and/or any role in a systematic organization (Shoemaker, 1975; Block and Fodor, 1972; but see Dennett, forthcoming). It would follow that they do not accord with the Cognitivist notion of a mental state or process.

But without even taking sides on the particular issue, I think we can see that it doesn't matter much to Cognitivism – which is, after all, only a theory of cognitive states and processes. In other words, if felt qualities are fundamentally different, so be it; explaining them is somebody else's business. This amounts to a kind of "segregation" of psychological phenomena, along roughly traditional lines. Such segregation can be legitimate (not a fudge) given one important assumption: Segregated noncognitive states can be effective in determining intelligent behavior

only insofar as they somehow generate quasilinguistic representations ("red there now," "left foot hurts") which can be accepted as inputs by the cognitive IPS. This assumption is plausible enough for felt qualities, and perhaps for some other states as well. I have in mind the much disputed "mental images" (Shepard and Metzler, 1971; Pylyshyn, 1973, in press; Paivio, 1975; Kosslyn and Pomerantz, 1977; Dennett, forthcoming). Since any Cognitivist theory must include some mechanism for getting from retinal images to cognitive descriptions of what is seen, I don't see why that same mechanism couldn't also take inputs from some precognitive visual "tape recorder" (perhaps one with "adjustments" for orientation, size, and location). Then playbacks from the recorder would have whatever nondiscursive, "image-y" quality perception has, and Cognitivism would be unruffled.

But I am much less sanguine about a similar segregation for moods. The difference is that moods are pervasive and all-encompassing in a way that felt qualities and images are not. The change from being cheerful to being melancholy is much more thorough and far-reaching than that from having a painless foot to having a foot that hurts. Not only does your foot seem different, but everything you encounter seems different. The whole world and everything in it, past, present, and future, becomes greyer, duller, less livable. Minor irritations and failings are more conspicuous and less remediable; ordinary things are no longer fun, lovely, or pleasing. If melancholy were an input representation ("melancholy here now") it would have to accompany and infect every other input, and transform the meanings of them all. But moods not only affect how things look, they affect how one thinks. What seems reasonable when you're cheerful seems foolish when you're melancholy, and *vice versa*. Likelihoods and improbabilities invert, as do what seems relevant to an issue and what seems beside the point.

Moods come upon us, but they are neither direct observations nor inferences. Many things affect our moods, but our moods also affect how things affect us; and in neither case is it quasilinguistic or rational. We do not state or believe our moods, or justify them on the basis of evidence or goals; they are just the way things are. In sum: Moods permeate and affect all kinds of cognitive states and processes, and yet, on the fact of it, they don't seem at all cognitive themselves. That suggests, at least until someone shows otherwise, that moods can neither be segregated from the explanation of cognition, nor incorporated in a Cognitivist explanation.

The second hurdle I want to mention concerns skills. I see three *prima facie* (not conclusive) reasons for doubting that the etiology of skillful behavior is cognitive. First, with rare exceptions, articulateness about a skill, no matter how detailed nor in what specialized quasilinguistic notation, is neither necessary nor sufficient for having it; it always takes practice, and often expert examples and talent ($\neq$ intelligence). Even a Rhodes scholar could not learn to play good ping-pong just from listening to thousands of detailed lectures about it; and even a Rhodes scholar ping-pong champion might be hard pressed to give a single detailed lecture on the subject. Second, a person who is acquiring or upgrading a skill may deliberately and thoughtfully try to execute certain maneuvers, but the thought and deliberation cease at just about the time the maneuvers become skillful and "natural"; the expert doesn't have to think about it. Third, skillful activity is faster than thought. Not only do skilled typists and pianists not have to think about what they're doing with their fingers; they can't. If they turn their attention to their fingers, as a novice must, their performance slows down and becomes clumsy, rather like a novice's.

A Cognitivist can explain these phenomena away by postulating some "unconscious" information processing which is somehow more efficient than, and immiscible with, that conscious thinking which is archetypically cognitive. But Dreyfus asks an interesting pointed question about this ploy, in the special case of chess skills (Dreyfus, 1972, p. 18). It is known that intermediate, advanced, and great chess players are alike in consciously considering on the order of a hundred plays in thinking out a move; they differ in their "skill in problem conception" (de Groot, 1965) – i.e., in preselecting which moves to think about. Now the rationales for these good preselections would be enormously long if they were spelled out (many thousands of plays). It's possible that players have some marvelously efficient unconscious information processor which works through these rationales; but if so, then why would anyone with such a splendid unconscious ever bother to deliberate consciously and tediously over a hundred plays? The implication is that the skillful preselection and the tedious cogitation differ not just in efficiency and consciousness, but in kind, and that neither could adequately substitute for the other. I think it would take powerful arguments (or prejudices) to outweigh this natural construal of the evidence – and only slightly less so in the case of skills in general.

But so what? If skillful behavior has to be explained in some non-Cognitivist way – call it "X" (maybe something to do with holograms) – then why not employ the segregation strategy introduced above for felt qualities and images? I think the danger here is not that the segregation strategy wouldn't work, but that it might work too well. "Skill" is such a broad and versatile notion that all kinds of things might fall under it. For example, the ability to act appropriately and adroitly in various social situations is a sort of skill, as is the art of conversation, and even everyday pattern recognition; moreover, these are like our earlier examples in that, to whatever degree one has mastered the skill, one needn't think about it to exercise it. But if very many such things turned out to be explainable in way X, rather than as the abilities of an IPS, then cognitive psychology would narrow dramatically in scope and interest. In the worst case, little would remain to call "cognitive" except conscious deliberation and reasoning – and that's hardly news.

The third hurdle I want to raise for Cognitivism is understanding; but this needs immediate qualification. In one sense, IPSs undoubtedly can understand, because computers programmed to be IPSs can do it. We could build a chess player, for example, that "understands" entered moves in any of three notations. What that means is that it responds appropriately (sensibly) to inputs in any of those forms. This is the same sense in which existing programs "understand" selected English sentences about colored blocks (see e.g., Winograd, 1972; and compare Greeno, 1977), airline reservations, and what not. Such usage is perfectly legitimate, but it's not all there is to understanding.

There is another notion of understanding, which, for convenience, I will call "insight" into why certain responses make sense, or are reasonable. As any teacher of arithmetic or logic knows, many students can learn the routines for getting the right answers, without the slightest insight into what's going on. And when original scientists struggle to find new and better threories, they grope for new "insights" into the phenomena, or new accounts which "make sense." Whether or not a new account, perhaps expressed in an unprecedented formulation, makes sense or is intelligible, is something which great scientists (and then their colleagues) can "just tell." Of course, whether an account is scientifically acceptable also depends on how well it accords with observations; but that does not determine whether it makes sense in the first place – both are necessary in science. The ability to tell when a whole account, a whole way of putting things, makes sense, is what I mean by insight.

The intelligibility of the whole account (or way of talking) then determines which particular utterances make sense, and what sense they make. Thus, it is only because quantum mechanics is an intelligible theory that one can make sense of talking about the wavelength of a particular electron (but not about the rest mass of a photon). And this brings us back to the conditions on interpreting something as an IBB. The testable requirement is that individual outputs make sense in the context of prior inputs and other outputs. But what determines which outputs would and

would not make sense in which contexts? That is, what determines which overall patterns render their constituents intelligible under an interpretation, or which input/output constraints count as cogency conditions? I have said that in appropriate circumstances, people can "just tell"; they can come to understand insightfully.

This is not to say that insight is itself some "transcendental" or impenetrable mystery, which we are forever barred from explaining. But once we appreciate that it is a genuine problem we can ask whether an IPS explanation could account for it. Now, we can understand how an IPS comes up with the reasonable outputs that it does, because we know how it works; in particular, we know that it works through a rationale for each output, and we know that it makes sense to say this of it because each of its interacting component IBBs consistently accords with certain cogency conditions. If we did not have that kind of a story to tell, then we would have no IPS explanation of the overall IBB's abilities.

So if an IPS explanation is to account for an object's having insight, then there must be a rationale for the insightful outputs. More specifically, if the insight is that certain new constraints constitute a kind of cogency, then there must be a rationale, according to the kind of cogency that the object and its components already exhibit, for why the new conditions count as cogency conditions. It seems to me that there could be such a rationale only if the new conditions were equivalent to, or a special case of, the established ones. For example, there could be no rationale according to chess player cogency conditions for why adding machine outputs make sense, or *vice versa*. If this is right, then an IPS with general insight into what makes sense would itself have to operate according to some cogency conditions that are ultimately general (so that the others which it recognizes could be given rationales as special cases).

There are two reasons to doubt that human insight can be explained that way. First, there is a sense in which it would preclude any radically new ways of understanding things; all new developments would have to be specializations of the antecedent general conditions. But I think the invention, say, of derivational-nomological explanation (around the time of Galileo) did comprise a "radical" advance in ways of understanding, in just the sense that the cogency of the new accounts could not be defended with a rationale which was cogent by prior standards. Medieval Aristotelians had explained (and understood) the motions of various kinds of bodies in terms of their efforts to get where they belonged, and their thwarting of each others' efforts. Galileo, Kepler, Newton, et al., didn't simply add to or modify those views. They invented a totally new way of talking about what happens, and a new way of rendering it intelligible; mathematical relationships and operations defined on universal measurable parameters became the illuminating considerations, rather than the goals and strivings of earth, air, fire, and water. I don't think a medieval IPS could have come to understand the new theory unless it had had it latently "built-in" all along. The same would be true of every IPS child who comes eventually to understand science, the arts, politics, and so on.

The second doubt has to do with this latent building-in – essentially, the ultimate general cogency conditions. We really have no reason to believe that there is any final characterization of what it is to make sense, except that it would facilitate a tidy account of intelligence. Barrels of philosophical ink have been spilt in the search for it, but so far without success. People who regularly make convergent decisions about the reasonableness of theories and interpretations don't explicitly work through rationales for their judgments. So we're back to postulating some mysterious and magnificent unconscious IPS. But once we admit that the phenomenon of insight is simply mysterious and unexplainable at present, then all we have to go on are the *prima facie* indications that IPS explanation is inadequate to the task.

In this section I have raised three issues which it seems to me Cognitivists must face, and which it is not yet clear they can handle. It is of course possible that successful treatments will eventually be found. On the other hand, if the approach is doomed to failure, I suspect that these are tips of some of the icebergs on which it will founder.

8. The state of the art

Needless to say, the eventual fate of cognitive psychology will be settled empirically – not by "armchair philosophizing." But the way in which experimental results bear on scientific theories, let alone whole approaches to the form that such theories should take, is seldom straightforward. In this concluding section, I will venture a few general points about Cognitivism and its relation to empirical observations.

It is illustrative to begin with cognitive simulation, a subdiscipline where cognitive psychology overlaps with artificial intelligence. A generation ago, the prospect of building intelligent computers inspired a lot of enthusiasm and brilliant work; but everyone must agree that results to date fall well short of early expectations. General problem solving programs have long since hit a plateau. Mechanical language translation has proven so elusive and frustrating that even military funding has dwindled. Advances in pattern recognition are painfully small, and still confined to contrived special "universes." Even game playing, a relative bright spot, is a disappointment against once confident hopes and predictions. About the only thing which exceeds original forecasts is the amount of computing power which has become available – and yet isn't enough. Does all this constitute an empirical refutation of the possibility of artificial intelligence? Not at all.

Perhaps the lesson is just that the problem was initially underestimated; soberer judges are now gratified by smaller steps in a longer trek, and disillusioned pessimists may still be exposed as carpenters who blamed their tools. On the other hand, if there were indeed something fundamentally misguided about the whole project, then recurrent bottlenecks and modest sparse successes are just what you would expect. The empirical record is simply ambiguous, and the real problem is to wrest from it whatever moral it does hold, as clearly and as helpfully as possible.

Cognitive simulation is not merely an incidental offshoot of cognitive psychology. It is a powerful and important research tool, because it provides a new and unprecedented empirical testing ground. Any IPS, or at least any one which is reducible to some level or dimension on which component input/output functions are expressible mathematically, can in principle be simulated on a computer. That means that simulations can function as concrete checks on whether particular proposed IPSs in fact have the abilities which they are supposed to explain. This is valuable when the proposed explanations are so complex that it is otherwise practically impossible to determine whether the things would actually work as claimed. In effect, the computer makes it feasible for Cognitivist theories to be more intricate and complicated than their predecessors could be in the past, and still remain under detailed empirical control.

By the same token, however, computer simulation serves as the front line where fundamental difficulties not resolvable by further complication would first show themselves. This is not to say that psychological experiments, and programmatic theories formulated with their guidance, are beside the point; quite the contrary, they form an essential high-level ingredient in the whole endeavor. But if one were genuinely to entertain the hypothesis that Cognitivism is misconceived, then the stumbling blocks empirically discovered by cognitive simulationists would be the first place to look for clues as to what went wrong.[4] How else than by struggling to build chess players could we have found out so definitively that the skill of deciding which moves to consider is not a simple matter of a few readily ascertained heuristics? What laboratory experiment could have

shown more clearly than the mechanical translation effort that the hardest thing to account for in linguistic performance is understanding what the discourse is all about?

If Cognitivism proves to be the wrong approach after all (that's still a big "if," of course), then the genius who makes the next basic breakthrough in psychology will probably take his or her cue from difficulties like these. Empirical indications of what cannot be done often pave the way for major scientific progress; think of efforts to weigh phlogiston, to build a perpetual motion machine, or to measure the speed of the Earth through the luminiferous aether.

A sense of history can give us perspective in another way. Until the rise of Cognitivism, Behaviorism reigned almost unchallenged in American psychology departments. It could boast established experimental methods, mountains of well-confirmed and universally accepted results, specialty journals carrying detailed technical reports, texts and curricula for teaching people to read and write those reports, and a coherent "philosophy" within which it all fit together and seemed inevitably right. In short, it had all the institutional earmarks of an advanced and thriving science. In retrospect, however, Behaviorism seems to have made little positive contribution to our understanding of the human psyche, and to be hopelessly inadequate to the task.

Kuhn's notion of a scientific paradigm can be extended in a way that sheds light on a situation like this (Kuhn, 1962). A *paradigm* is a major scientific triumph, so impressive in breaking new ground, and yet so pregnant with unfulfilled possibilities, that a technical research tradition coalesces around it as a model. Thus, the achievements of Thorndike and Pavlov inspired a vigorous and sophisticated investigation of the conditioning of birds, dogs, and rats – and also of people, to the extent that they are similar. But most of the interesting and important aspects of intelligent behavior, exhibited especially by humans, turn out to involve processes qualitatively different from those discovered by Thorndike, Pavlov, and their followers. So when Behaviorism was taken as an approach to psychology in general, its paradigm became a kind of impostor; experiments, concepts, and methods which were genuinely illuminating in a limited domain posed as the model for illumination in a quite different domain, where they had virtually no demonstrated credentials, and really didn't belong.

Cognitivism is a natural development from Behaviorism. It retains the same commitment to publicly observable and verifiable data, the same rejection of posits and postulates that cannot be treated experimentally, and the same ideal of psychology as a natural science. Its advantage is having shown, via the systematicity and intentional interpretation "cornerstones," how to make good empirical sense of meaningful or rational internal processes – which gives it a much richer and more powerful explanatory framework. And not surprisingly, it has now acquired the institutional earmarks of an advanced and thriving science. But cognitive psychology too can be accused of having an impostor paradigm. The concrete achievements which inspire the notion of IPS explanation, and prove it to have application in the real world, come originally and almost entirely from the fields of computer science and automatic data processing. The few cases in which people explicitly and deliberately work through a rationale do suggest an analogy; but so did the cases in which people responded to conditioning.

Like their predecessors, Cognitivists have made undeniably important and lasting discoveries. But also as before, these discoveries are conspicuously narrow, even small, compared to the depth and scope of psychology's pretheoretic purview. The brilliance of what has been done can blind us to the darkness that surrounds it, and it is worth recalling how many shadows Cognitivism has not (yet) illuminated. How is it, for example, that we recognize familiar faces, let alone the lives reflected in them, or the greatest of Rembrandt's portrayals? How do we understand conversational English, let alone metaphors, jokes, Aristotle, or Albee? What is common sense, let alone creativity, wit, or good taste? What happens when we fall asleep, let alone fall under a spell, fall apart, or fall in love? What are personality and character, let alone identity crises, schizophrenia, the experience of enlightenment, or moral integrity? We turn to psychology if we think these questions have scientific answers; and if we shouldn't, why shouldn't we? Cognitivists are as vague and impressionistic on such issues as psychological theorists have always been. Of course, they too can buy time with the old refrain: "be patient, we're only just beginning (though so-and-so's preliminary results are already encouraging)." Promissory notes are legitimate currency in vigorous sciences, but too much deficit spending only fuels inflation.

The human spirit is its own greatest mystery. Perhaps the idea of an information processing system is at last the key to unlocking it; or perhaps the programmable computer is as shallow an analogy as the trainable pigeon – the conditional branch as sterile as the conditioned reflex. There is no way to tell yet, but we should be as ready to follow up on partial failures as we are on partial successes. The clues could be anywhere.

ACKNOWLEDGMENTS

In preparing this paper I have incurred more of a debt than I can properly express to the inspiration and constant guiding criticism of H. L. Dreyfus. I am also thankful to several students and colleagues for directing my attention to weaknesses in earlier drafts (some of which, no doubt, remain) – especially Bob Brandom, Dan Dennett, Jay Garfield, Allan Gibbard, Bill House, and Zenon Pylyshyn. Finally I am grateful to the University of Pittsburgh Faculty Grants Committee for research support during the summer of 1975.

NOTES

1. For readers familiar with the work of Quine, I would like to clear up what I think is a common misunderstanding. Quine is a Behaviorist of sorts, and he sometimes seems to defend that on the basis of his doctrine of the indeterminacy of translation (Quine, 1960, Ch. 2). Thus, it's natural to suppose that Cognitivism is as opposed to the latter doctrine as it is to Behaviorism. It isn't. In the terminology of this paper, Quine's claim is the following: For any IBB, there are many different intentional interpretations of the same input/output typologies, which are all equally "good" by any empirical tests; that is, they are all such that the outputs consistently make reasonable sense in context. Hence, one's "translation" of the inputs and outputs is empirically indeterminate, at least among these options. Now, it might seem that if the IBB were an IPS, and if one knew what it was "thinking" (its internal cognitive processes), then one could determine what its outputs really meant, and thereby undercut the indeterminacy. But if Quine is right in his original claim (and I take no stand on that), then it applies to the interpretations of the component IBBs as well. Thus the indeterminacy, rather than being undercut, is just carried inward; in Quine's terms, all the translations are "relative to a translation manual." That would no more rule out Cognitivism than it would linguistics.

2. Strictly, the required relation between the two sets of constraints is weaker than isomorphism. It suffices if every input/output pattern which would satisfy the explained constraints on the lower dimension would also satisfy the cogency conditions on the interpretation being reduced (the constraints on the upper dimension). This amounts to saying that the instantiation can explain more than the IBB ability in question – e.g., not only how it manages to play chess, but also why it always neglects certain options.

3. To my knowledge, Dennett has made the fullest previous attempts to explicate intentional reduction, but they seem to me incomplete. In an early paper (Dennett, 1971), he distinguished the "intentional stance" (having to do with interpretation and meaning) and the "design stance" (having to do with systems engineering). But his examples suggest that in the design stance one attends to the instantiating physical system, while in the intentional stance one attends to the overall IBB – leaving out entirely the crucial notion of an IPS. In a later paper (Dennett, 1975), he effectively remedies this by introducing "committees of homunculii" (essentially bureaucracies of IBBs; see my metaphor in section 1). Then by suggesting that each homunculus should be analyzed as a committee of stupider homunculii, until the stupidest can be "replaced by machines," he has what amounts to an intentional reduction. But he still lacks the distinction of dimension and level, and hence the notion of intentional instantiation.

4. One of the special merits of Dreyfus' "critique of artificial reason" (Dreyfus, 1972) is his attempt to gain some positive insights from the tangle of frustrations that have bedeviled cognitive simulation. See his pioneering effort (Ch. 1) to extract some order from the apparent chaos, via the four distinctions: fringe consciousness vs. heuristically guided search; ambiguity tolerance vs. context-free precision; essential/inessential discrimination vs. trial and error search; and perspicuous grouping vs. character lists.

REFERENCES

Baron, R. J. A Model for Cortical Memory. *Journal of Mathematical Psychology*, 1970, 7, 37–59.

Block, N., and Fodor, J. What Psychological States are Not. *Philosophical Review*, 1972, 81, 159–81.

Campbell, F. S. The Transmission of Spatial Information through the Visual System. (In: Schmitt and Worden, eds., 1974).

Cathey, W. T. *Optical Information Processing and Holography*. New York: John Wiley & Sons, 1974.

Cavanagh, J. P. Holographic Processes Realizable in the Neural Realm. Unpublished doctoral dissertation, Carnegie Mellon University, 1972.

Cummins, R. Functional Analysis. *Journal of Philosophy*, 1975, 72, 741–65.

Davidson, D. Mental Events. In: Foster and Swanson, eds. *Experience and Theory*. Univ. Mass. Press, 1970.

Radical Interpretation. *Dialectica*, 1973, 27, 313–28.

de Groot, A. *Thought and Choice in Chess*. The Hague: Mouton, 1965.

Dennett, D. Intentional Systems. *Journal of Philosophy*, 1971, 68, 87–106.

Why the Law of Effect Will Not Go Away. *Journal for the Theory of Social Behavior*, 1975, 5, 169–87.

Why You Can't Make a Computer that Feels Pain. *Synthese* (forthcoming).

On the Absence of Phenomenology. (unpublished).

Dreyfus, H. L. *What Computers Can't Do*. New York: Harper and Row, 1972.

Erickson, R. P. Parallel "Population" Neural Coding in Feature Extraction. (In: Schmitt and Worden, eds., 1974).

Firth, I. M. *Holography and Computer Generated Holograms*. London: Mills & Boon, 1972.

Fodor, J. Explanation in Psychology. In: M. Black, ed., *Philosophy in America*. Ithaca: Cornell University Press, 1965.

Special Sciences (or: The Disunity of Science as a Working Hypothesis). *Synthese*, 1974, 28, 97–115.

Frey, P., and Adesman, P. Recall Memory for Visually Presented Chess Positions. *Memory and Cognition*, 1976, 4, 541–47.

Gabor, D. Associative Holographic Memories. *IBM Journal of Research and Development*, 1969, 13, 156–59.

Gödel, K. Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte für Mathematik und Physik*, 1931, 38, 173–98.

Goodman, N. *Languages of Art*. Indianapolis: Bobbs–Merrill, 1968.

Grandy, R. Reference, Meaning and Belief. *Journal of Philosophy*, 1973, 70, 439–52.

Greeno, J. Process of Understanding in Problem Solving. In: Castellan, Pisoni, and Potts, eds., *Cognitive Theory II*. Hillsdale, N.J.: Lawrence Erlbaum, 1977.

Harman, G. *Thought*. Princeton: Princeton University Press, 1973.

Hearst, E. Psychology across the Chessboard. *Psychology Today*, June 1967.

Hempel, C. G., and Oppenheim, P. Studies in the Logic of Explanation. *Philosophy of Science*, 1948, 15, 135–75.

Herriott, D. R. Applications of Laser Light. *Scientific American*, September 1968.

Hudson, L. *The Cult of the Fact*. London: Cape, 1972.

Kabrisky, M. *A Proposed Model for Visual Information Processing in the Human Brain*. Urbana: University of Illinois Press, 1966.

Kosslyn, S. M., and Pomerantz, J. R. Imagery, Propositions, and the Form of Internal Representations. *Cognitive Psychology*, 1977, 9, 52–76.

Kuhn, T. *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press, 1962 (revised ed. 1970).

Leith, E. N., and Upatnieks, J. Photography by Laser. *Scientific American*, June, 1965.

Lewis, D. Radical Interpretation. *Synthese*, 1974, 23, 331–44.

Lucas, J. R. Minds, Machines and Gödel. In: A. Anderson, ed., *Minds and Machines*. Englewood Cliffs, N.J.: Prentice–Hall, 1964.

Minsky, M., and Papert, S. Artificial Intelligence Progress Report. Artificial Intelligence Memo No. 252, MIT, 1972.

Nagel, E., and Newman, J. R. *Gödel's Proof*. New York University Press, 1958.

Paivio, A. Imagery and Synchronic Thinking. *Canadian Psychological Review*, 1975, 16, 147–63.

Pollen, D. A., and Taylor, J. H. The Striate Cortex and the Spatial Analysis of Visual Space. (In: Schmitt and Worden, eds., 1974).

Pribram, K. H. *Languages of the Brain*. Englewood Cliffs, N.J.: Prentice–Hall, 1971.

How is it that Sensing So Much We Can Do So Little? (In: Schmitt and Worden, eds., 1974).

Pribram, K. H., Nuwer, M., and Baron, R. J. The Holographic Hypothesis of Memory Structure in Brain Function and Perception. In: D. H. Krantz, et al., eds., *Contemporary Developments in Mathematical Psychology, II*. San Francisco: Freeman, 1974.

Putnam, H. Minds and Machines. In: A. Anderson, ed., *Minds and Machines*. Englewood Cliffs, N.J.: Prentice–Hall, 1964.

Mind, Language and Reality. Cambridge: Cambridge University Press, 1975.

Pylyshyn, Z. What the Mind's Eye Tells the Mind's Brain: A Critique of Mental Imagery. *Psychological Bulletin*, 1973, 80, 1–24.

Imagery and Artificial Intelligence. To appear in *Minnesota Studies in Philosophy of Science*, IX. (in press).

Quine, W. V. O. *Word and Object*. Cambridge, Mass.: MIT Press, 1960.

Schmitt, F. O., and Worden, F. G., eds. *The Neurosciences: Third Study Program*. Cambridge, Mass.: MIT Press, 1974.

Sellars, W. *Science, Perception and Reality*. London: Routledge & Kegan Paul, 1963.

Shepard, R. N., and Metzler, J. Mental Rotation of Three-Dimensional Objects. *Science*, 1971, 171, 701–703.

Shoemaker, S. Functionalism and Qualia. *Philosophical Studies*, 1975, 27, 291–315.

Simon, H. A. *The Sciences of the Artificial*. Cambridge, Mass.: MIT Press, 1969.

van Heerden, P. J. A New Method of Storing and Retrieving Information. *Applied Optics*, 1963, 2, 387–92.

Wilson, N. L. Substances without Substrata. *Review of Metaphysics*, 1959, 12, 521–39.

Winograd, T. Understanding Natural Languages. *Cognitive Psychology*, 1972, 3, 1–191.

Yevick, M. L. Holographic or Fourier Logic. *Pattern Recognition*, 1975, 7, 197–213.

Open Peer Commentary

Commentaries submitted by the qualified professional readership of this journal will be considered for publication in the Continuing Commentary sections of forthcoming issues.

Note: Commentary reference lists omit works already cited in the target article (as indicated by op. cit.)

by **John H. Andreae**

Department of Electrical Engineering, University of Canterbury, Christchurch, New Zealand.

On inference from input/output. My only justifications for commenting on the abstruse paper of Haugeland are, first, that it is a basic aim of this journal to get people to tread in each other's territories, and, second, that the paper is, presumably, trying to help those, like myself, who have tentative working models of the mind.

Stripped of its philosophical dressing, the paper seems to be offering as a central concept the IBB, or meaningful black box. The IBB is characterized by its outputs' making "reasonable sense" in terms of its histories of inputs. According to Haugeland, this reasonable sense "is seldom hard to recognize in practice," although he gives no black box or cognitive examples of its truth. Yet, in the case of quite simple black boxes, whose structures and functions are really unknown (i.e., black), there are tremendous difficulties in deriving structural models that make reasonable sense of behaviors. Quite small amounts of memory and indeterminacy generate great difficulties in going from behavioral descriptions to structural descriptions.

The human brain, at the present time, is still a very black box, and we cannot afford to neglect any direct structural clues that can be obtained from

neuropsychology and neurophysiology. At the cognitive level, even the nature of the inputs and the outputs of the system are in question. In synthesizing a model of the mind, we have to work up from the structural information and down from the behavioral information, making inspired guesses as to how they may be related. Simplification, the essence of understanding, can result both from reduction to structure and from reduction to behavior. As our models improve, vague terms like mood, feeling, insight, intuition, understanding, intelligence, and so on, will be replaced by more definite ones. In so far as Cognitivism recognizes the need to take into account the states of a system as well as its input/output behavior, few would exchange it for behaviorism. However, Haugeland seems to be suggesting by his IBB that we can test the validity of some proposed state-transition structure solely by means of input/output behavior (i.e., input from the environment and output to the environment). Common experience with quite simple systems (like someone else's electronic circuit!) indicates that this is not so.

Perhaps I am saying no more than that it is unreasonable to expect to be able to define "a specified range of phenomena" in which sets of input types and output types are "uniquely determinable" in the way required by Haugeland for his first step in constructing an IBB. The weakness of his behaviorist assumption that everything can be derived from input and output is well illustrated by his "hurdles" of mood, skill, and insight. So long as one neglects to use knowledge of internal states (derived from physical and chemical measurements and controls), as in Haugeland's construction of an IBB, subconscious activity of the mind, which accompanies everything we do, is bound to invalidate our explanations. What remains when moods, skills and insight are left out?

by **Michael A. Arbib**

Computer and Information Science Department, University of Massachusetts, Amherst, Mass. 01003

On making distinctions that are not maintained. Haugeland's paper appears to be a textbook example of the sort that convinces scientists (mistakenly, I would insist) to conclude that philosophers have nothing of value to say to them. The author intends to give a critique of cognitivism in psychology and philosophy. He does not discuss even one example of a cognitivist analysis of some intelligent behavior – the closest he comes is the observation that programming a computer-based chess player in LISP is a nonphysical instantiation. His one example of a noncognitivist analysis is a fanciful "holographic chess-player" – hardly reassuring to the scientist since (a) it ignores the strategic elements of the game and speaks only of recognizing situations; and (b), with some eminent exceptions, most neuroscientists regard the holographic "model" of the brain as being, at best, only a weak metaphor. Sections 1 to 5 are devoted to the making of distinctions that are little used in the critique proper (sections 6, 7, and 8), that ride roughshod over the reality of current scientific methodology and that are not equipped with proper delimiting criteria. All that one needs from these sections is an understanding of Haugeland's slogan that "Cognitivism is the position that mind is to be understood as an IPS." Before explaining this, let me devote four paragraphs to these distinctions that Haugeland makes but does not/cannot maintain.

Section 1 offers a sharp distinction between *derivational-nomological* explanation (Newton's laws), *morphological* explanation (replication of DNA), and *systematic* explanation (explaining the behavior of a system in terms of its functional components). What I object to is the suggestion that these are almost exclusive styles of explanation, when, in fact, the three can occur in virtually any combination. Using Newton's laws to analyze a ball rolling on a surface requires morphological constraints; analysis of DNA conformation involves physics as much as bell-and-stick chemistry; and a system theorist analyzing an ecological system may well use differential equations to predict system behavior. Haugeland admits the distinctions may be somewhat blurred, but makes them so that he can restrict cognitivism to systemic explanation. We shall return to this point (good for philosophy, bad for science?) below.

Section 2 argues that a scientifically respectable reduction need not be a reduction to physics modelled on the statistical reduction of thermodynamics to mechanics. As a computer scientist, I certainly agree – a program's function is independent of the physical implementation of the machine language functions. This is probably the argument that satisfies many cognitive scientists. However, as a neuroscientist I must take issue with this form of of cognitivism, for the style of certain types of human cognition may well depend on the fact that they are implemented in a brain made of components susceptible to the action of hormones, and there may be no satisfactory reduction that does not take this chemical activity (that is, the fine details of circuit-level implementation) into account.

Section 3 labors mightily to define an IBB (*intentional black box*). This is simply a black box in which the inputs and outputs can be described in some symbolic notation that a human being can interpret in such a way that "it is shown empirically that under the interpretations the actual outputs consistently make reasonable sense in the context (pattern) of actual prior inputs and other actual outputs." The author does not know of any criteria for deciding when a BB is I, shrugging it off with ". . . it doesn't matter much in actual field or laboratory work, because by and large everyone can agree on what does and doesn't make sense. . . . If one is knee-jerk liberal about what makes sense then all kinds of objects can be trivially interpreted as IBBs . . . [but] I will assume that such cases can be ignored." One hopes that this "Y'know what I mean?" style of definition will not catch on. The upshot of the section is what every system theorist knows – a respectable theory can (and usually does) assign interpretations to the inputs and outputs of a system. In fact, the judicious choice of such an interpretation is usually the first step in successful systems analysis.

Section 4, recalling that a systematic explanation shows an ability as resulting from the organized cooperative interactions of various functional components, specifies further that an IBB is to be called an IPS (*information processing system*) if it can be systematically explained without "deinterpreting it; for example in a chess program, each functional component must be doing something chess-like (e.g., exploring moves) rather than nonchess-like (e.g., printing out marks that can be interpreted as signalling a move). Section 5 notes that repeating such reductions must eventually lead to components for which in further reduction deinterpretation *must* take place – Haugeland calls this *instantiation*. This instantiation is *physical* when the functional components are expressed in physical terms; *intentional* when the functional components are IBBs involving a different subject matter. Note well, then, that a typical computer program is *not* an information processing system in Haugeland's sense since it involves deinterpretation into the level of the programming language. One hopes that this "inconsistent with normal usage" style of terminology will not catch on.

Back to Haugeland's summary of cognitivism in the slogan "the mind is to be understood as an IPS." The translation is as follows: "Cognitivism analyzes the mind as a system composed of functional components in which the inputs and outputs of the system itself and of these components fall under a common intentional interpretation. Further reduction may be possible under this interpretation, or may require new interpretations or the transition to a physical level of description."

Section 6 correctly notes that the structural composition of the brain as a network of neurons does not guarantee that these neurons are organized into subsystems that admit intentional interpretation (though, of course, the possibility that they do provides a variety of subgoals for the neurosciences). It then offers the hologram metaphor to show that a system need not proceed by explicitly working out the rationale for its actions. The defects of holography as a brain model do not vitiate its use as a debating point, but I do want to note one apparent weakness in Haugeland's analysis. He talks as if a *functional* component must also be a *structural* component. An associative memory might well be analyzed into subsystems suited to recognize positions of pieces, special configurations, and sequences of moves – and so be an IPS – even if these functions are implemented in overlapping networks. The deinterpretation need only occur at another level down the reduction hierarchy.

At this stage I must make a confession of ignorance. I do not know what "cognitivism" is. I do know what might be called "cognitive simulation" – the use of programs modelled on those of artificial intelligence to explain phenomena in cognitive psychology (See Pylyshyn et al., *BBS* 1:1). I believe that this approach is a fruitful one, but that evolutionary and pharmacological considerations must be used to enrich it, as must the "derivational-nomological" techniques of mathematical system theory. Thus in approaching the final section of Haugeland's paper, I must look at three questions. (1) Since I do not know what cognitivism is, does it equal "cognitive simulation?" (2) Are Haugeland's discussions of limitations of cognitivism *philosophically* compelling? (3) Are there *scientific* reasons to reject pure cognitivism?

In section 8, Haugeland talks as if cognitive simulation were what his paper was about. Yet I wonder if his notion of an IPS really squares with work in that area. In Samuel's (1959, 1967) checkers-playing program, the adjustment of weights is a crucial feature of the learning procedure and seems to provide a successful model of skill acquisition. Yet, in section 7, Haugeland spells out his doubts that "the etiology of skillful behavior is cognitive," listing this as one of three hurdles for cognitivism. Now I agree that skillful behavior is not cognitive in Haugeland's sense, but I argue that Samuel's checker-player is (or can be the basis for) an excellent example of cognitive simulation. Thus Haugeland's notion of cognitivism is too impoverished to include cognitive simulation.

Haugeland's argument against cognitivism in section 7 is that moods, skills, and understanding are hard to implement as IPSs. In section 8 he argues that cognitive simulation has reached a plateau. But, in fact, A.I. is making great progress with "understanding," and cognitive simulation, if expanded to include identification/parameter adjustment methods, can address many problems of skills. It may well be true that Haugeland-sense cognitivism cannot, though I am not sure that his "I don't see how to do it" arguments carry much philosophical weight.

This brings me to the close. The philosopher may ask "Can cognitivism explain the mind?" and the scientist may ask "Should I restrict myself to cognitivism in explaining the mind?" I am all for scientific pluralism, believing that we poor mortals must specialize in our choice of problems and tools to solve them. I do not think many cognitive psychologists would choose to be cognitivists in Haugeland's sense, but many will choose to use cognitive simulation as an explanatory tool. Some will choose to ignore system theory and neuroscience (*inter alia*) in their own work, others will insist that these are irrelevant to all of cognitive psychology. I understand the strategic decision of the former and reject the dogmatism of the latter. Accepting Haugeland's examples, I expect that neuropharmacology will play a central role in the analysis of mood, and that biomechanics and control theory will be coupled with the theory of adaptive networks in the analysis of skills. But for the study of understanding, I think cognitive simulation is the current best bet. These are not philosophical judgments, but a personal opinion on allocation of scientific resources.

REFERENCES

Samuel, A. L. Some Studies in Machine Learning Using the Game of Checkers. *IBM Journal of Research and Development*. 3:210–229, 1959.

Some Studies of Machine Learning Using the Game of Checkers. II. Recent Progress. *IBM Journal of Research and Development* 11:601–617, 1967.

by **Margaret Atherton**

Department of Philosophy, Brooklyn College, City University of New York, Brooklyn, N.Y.

The scope of cognitivism. Defenses of cognitivism, that in the long run strike me as more harmful than helpful, have quite often embodied two general attitudes toward the theory being defended. The first is that cognitivism not only represents an alternative that is an advance over previously existing theories like behaviorism, but that it also constitutes a unique alternative, so that to reject cognitivism is to throw psychology back into the dark ages of behaviorism. The second attitude, which provides in some sense a reason for the first, is that the description of mental processes provided by cognitivism is the only way in which it is plausible to imagine a mind can work. Thus we must either describe the mind as an Information Processing System or give up hope of explaining how it works. What seems to me to be the strength of Haugeland's position is that he shows it is possible to accept the cognitivist's claim to offer a psychological theory that is an advance over behaviorism, while still suggesting that cognitivism may turn out to be a limited theory in the way behaviorism did. Thus, Haugeland encourages us to look in a positive light at the problems that exist for cognitivism as problems that should encourage us to go forward and not to go back.

The strength of cognitivism, as Haugeland states it, is that it legitimates an explanation of mental events and processes in mental terms. It encourages us to think that we can answer questions like "Why did the chicken cross the road?" in terms of thought, plans, hopes, desires, and other states and processes belonging to and ascribable to the chicken. Our explanation of its road-crossing will mention the various things that make this activity meaningful to the chicken. Such an account is scientifically respectable because, while the inputs to and the outputs from the chicken cannot tell the whole story of why it crossed the road, they nevertheless constitute good empirical evidence for the rest of the story by serving as evidence for the various mental events we ascribe to the chicken to fill in the gap between input and output. The weaknesses of the cognitivist story stem from the fact that cognitivism makes very stringent demands on what kind of mental life can appropriately be postulated to fill the gap. The assumption is that the behavior in question can be explained only if it can be accounted for as the result of some sort of information-processing or hypothesis-testing. In general, as Haugeland points out, the input must be assumed to be able to take the form of data posing a problem, and thus must assume some sort of quasi-linguistic nature, must be assumed to be able to take the form of a solution to the problem, and the postulated mental states must be assumed to be able to take the form of steps or calculations in the solution of the problem from the data. But the difficulty is that on this account every mental state must turn out to be some sort of proposition and every mental process some sort of calculation. Anything that does not immediately fit this model must either be assumed not to be mental, not to deserve a psychological explanation, or to be jammed willy-nilly into the model. And this either requires us to refuse to call some things mental that otherwise seem to deserve psychological explanation or else dictates that an explanation take what can seem to be a highly unnatural form.

I think Haugeland is quite right that moods and emotions are generally very much the kinds of things that we would like to be able to explain in mentalistic terms. Being in love is not explicable in terms of dispositions to approach the beloved object; being fearful is not just the disposition to avoid. As is the case with other mental states, moreover, we want to be able to attribute the emotion *to* the organism, to say the emotion belongs to it. Nor are moods quite so free from evidential support as Haugeland suggests. It has been pointed out that it is possible to argue about the appropriateness of a mood, to say, for example, "you ought not to be afraid of the big dog." But cognitivism is unwilling to talk about the presence of events in a mind except insofar as the events give rise to quasi-linguistic representations or to beliefs. Yet there seem to be distinctions between having a mood and beliefs about that mood, as indeed there are between things like being conscious of a scene or a throbbing in the temples and having beliefs about any of these things. Still, it seems hard to understand in what sense one is said to be having the mood, or the throbbing in the temples, or awareness of the scene without these things being represented to one, without their being present in some form or other in one's mind. Thus it is hard to understand in what sense one is thought to have a sensation in a prerepresentational form. On the other hand, it seems clear that no linguistic description can capture the scene or mood of which one is aware, or that the awareness can be identified with a set of beliefs about what one is aware of. Thus it seems to me that sensations have more in common with moods than Haugeland allows and that, in both cases, a decision not to count them within the domain of psychology is unsatisfying. It seems clear that things like sensations and moods do serve as input for us, just as do thoughts and beliefs, and so it seems to me not at all clear that it is correct to select just one mode of input as relevant to psychology, even though we have a way of describing thoughts and beliefs – quasi-linguistic representation – which is not so obviously available for moods and sensations.

The reason it is thought to be necessary to reduce all forms of input to quasi-linguistic representations, even in cases where such representations seem to fall far short of what is actually available (such as "left foot hurts" or, presumably, "pang of love here now", is that there otherwise seems to be no way to show how inputs can enter into the calculations that are thought to constitute the only possible description of mental processes. But, as Haugeland and many others have pointed out, requiring a description in terms of calculating results is assimilating all mental skills and process to that of deliberation and reasoning. And certainly many skills, from typing to chicken-sexing to chess-playing do not seem to involve deliberation and calculation. Thus it looks as though cognitivists have simply selected one kind of skill, deliberation, as previously they had slected one form of input, and have insisted that all other skills unconsciously conform to this model, when consciously they do not. But, of course, what the cognitivist wants us to recognize is that the apparent lack of calculation always stems from the fact that we have not as yet developed an explanation for how the skill is performed. This is because the cognitivist feels the explanation is incomplete without the calculation; we have not explained, for example, how faces get

recognized, until we have shown exactly what information about faces could be considered available and exactly what calculations are involved in reaching the conclusion that this is a face.

But it may well be that all skills necessarily appear to take the form of calculations, not because calculating is any more than any other method, obviously the way we do things, but because calculating has a significant place in the way we usually *explain* things. We do not think we have adequately justified some conclusion until we have explained the reasons for holding it. But there is no particular reason why the process of reaching a conclusion should be the same as the process of justifying that conclusion. And, moreover, even if we accepted the model of justifying a conclusion as the model to use in interferring unconscious mental states and processes, Haugeland shows in his discussion of understanding that what actually goes on when we seek to devise an explanation is not at all like the cut and dried process that is assumed to be taking place internally. Devising conditions of cogency for an explanation is quite often a part of developing an explanation, and conditions of cogency will change as discoveries shed new light on the nature of problems. Therefore, it seems extremely unlikely that there should be one unique method of constructing a chain of reasoning that is *the* method belonging to us, psychologically speaking, and coded in us to determine the course of our mental processes.

What is attractive about cognitivism is that it shows that it is not out of the question to develop theories about the mind's using mental terms, in which we can use our experience of minds to devise these theories. But if we pay attention to what actually goes on in one's mind and to the kinds of skills one possesses, there seems no particular reason to believe that the information processing model is the only nonbehaviorist model available.

by **Robert J. Baron**

Department of Computer Science, University of Iowa, Iowa City, Iowa 52242

On explanation, holograms, moods, and skills. Haugeland has done an admirable job of presenting the nature of cognitivism and the cognitive approach to understanding brain function. Although I agree with most of his points, there are a few to which I must take exception.

Haugeland first discusses three styles of scientific explanation: deductive-nomological (including derivational-nomological), morphological, and systematic. He claims that "only the systematic style is directly relevant to Cognitive Psychology." I disagree. Haugeland's major emphasis is concerned with the mathematical style employed by physicists ("Thus the charge of slavishly imitating mathematical physics does not apply to Cognitivism. . . ."), and not the discrete mathematics of computer science (automata theory, formal language theory, symbol manipulation systems, etc.). Computer scientists have long been concerned with computation theory, what can and cannot be computed, how fast various computations can be performed, and the correctness of the algorithms of computation. The computation theoretic style of explanation, when applied to computational systems, is deductive-nomological. Furthermore, the brain shares more of the features of a computational system (computer, automata, formal language, etc.) than it does of a physical system. In particular, cognitive psychologists talk of memory, information, computation, learning, and control just as do computer scientists. I therefore suggest that Haugeland should not rule out the deductive-nomological style of explanation of cognitive psychology. But I agree that "slavishly imitating mathematical physics" is not the thing to do.

Another recent formalism that appears to have great promise is the simulation formalism currently being developed by Rieger and Grinberg (1976). This formalism enables a precise description of complex systems and of interactions between system components, and it also appears to be the framework for deductive-nomological explanation of systematic behavior. If I am correct, then, proof techniques, to be developed and applied to such descriptions, may be a formalism that is badly needed in cognitive psychology.

In section 6 of his paper, Haugeland argues that there is in reality no proof that a cognitive approach to understanding brain function will ultimately succeed. He brings into his arguments a discussion of several "holographic" models of memory. I want to emphasize that the holographic models, along with other models for storage, are generally proposed as associative storage networks. These models account for some behavioral aspects of brain function (association [recognition] and forgetting the typical), but almost any model for associative storage would account for the same aspects of behavior. One should not think of the holographic models as any more than they were intended to be: plausible mechanism for storage, association, and recall of information. They are not general models of brain function: in fact, they are far from it. Behavioral properties that may be discussed by the model builder, such as learning, deduction, and forgetting, are not generally dependent on the fact that the storage process is holographic. In fact, many of the properties of optical holograms (three-dimensional reconstruction, preservation of color, resistance to damage, etc.) are not directly relevant to the neural models. In summary, one must study each model to determine exactly which properties are ascribed to the storage mechanism, and which properties to the control mechanisms of the storage system (when to store information, where to store it, whence to recall information, etc.). Otherwise, the confusion that results negates the usefulness of the model as an explanation for the particular function of interest. Haugeland's mention of the holographic models, followed by a discussion of a plausible move generator for a chess player, unfortunately distorts rather than clarifies the role of the holographic memory models.

Haugeland makes one point that merits further discussion. He states, "Optical holograms are two-dimensional and static, whereas the brain is three-dimensional and dynamic." In order to make any sense of such a statement, one must interpret "static," "dynamic," "two-dimensional," and "three-dimensional."An optical hologram is static: once photographically developed, the distribution of silver densities across the hologram is fixed. The same is most likely true of the permanent memory traces of experience in the brain: once the neural parameters, whatever they are, are established and fixed during consolidation, they become static. Although there is no direct proof that memory traces of experience are permanent, the work of Penfield and Roberts (1959) and Penfield and Perot (1963) certainly suggests that they are. When Haugeland uses the term "dynamic," he is either referring to the fact that information patterns in the brain are conveyed by *active* patterns of neural activity, or he is suggesting that there are no permanent memory traces in the brain. If the reference is to active patterns of neural activity, then there is a direct analogy with the optical systems. The light pattern that illuminates and interacts with the hologram to produce the reconstructed wavefront is dynamic. In each case, a static pattern (distribution of silver intensities, neural parameters) is used to convert a dynamic pattern (the incident light wave, a neural discharge pattern) into another dynamic pattern (the synthesized wavefront, the recalled neural firing pattern). Finally, if Haugeland uses the term "dynamics" to refer to the very complex way that information is routed and processed within the brain as an information processing system, then the comparison is unfair. The holographic aspect of the memory models was concerned only with the storage and synthesis processes of memory, not with the more complex control processes that must take place. In summary, the statement that holograms are static while the brain is dynamic does not make sense.

My next comment is on the concepts of two- and three-dimensional processing systems. There are both two- and three-dimensional holograms. Van Heerden's (1963) method of optical information storage is three-dimensional and holographic, although he did not use that terminology. Three-dimensional holograms have been used for storing moving images by continuously changing the angle of the reference wave during the storage and recall processes, and they have been used for storing multiple images in the same film. In contrast, there is no conclusive evidence that the brain is a three-dimensional information processing system analogous to the three-dimensional storage system of van Heerden. (Note that Beurle, 1956, one of the earliest neural modelers, viewed the brain as a three-dimensional storage medium and proposed a storage system analogous to van Heerden's.) Information in the brain progresses through layers of neurons, and each layer performs its specific transformation on the input pattern. It is just as likely that this is a sequence of two-dimensional transformations as it is a sequence of three-dimensional transformations. In summary, the importance of the three-dimensional nature of the brain is not yet known, and it may have very little to do with the particular process being performed.

My final comments relate to the serious hurdles proposed by Haugeland. I do not share his concern about the difficulty of discovering a cognitive (functional) basis for moods. Let me first posit that perception is dependent on past experiences and how they relate to current events. If the behavior of our memory stores (including the storage mechanism, the association mechanism, the recall mechanism, and the control mechanisms) is systematically modified, then our perceptions will also be altered. The state-dependent learning studies presented by Overton (1965) may be an extreme

experimental example of such chemically induced alterations. Modification that can result in such state-dependent phenemona can be induced in almost every current model of memory by systematically changing some of the coupling parameters between cells in the memory stores. In the model I proposed (1970 *op. cit.*) for information storage and association, for example, one need only reduce all excitatory coupling coefficients by a reasonable percent to see this effect. (Reducing both excitatory and inhibitory coupling coefficients proportionately does not produce the effect.) If moods are related to brain chemistry, and there is certainly reason to believe they are, and if brain chemistry is related to neural coupling parameters, then one would expect systematic modifications due to the chemistry of moods to alter our current perception.

The relationship between skills and other behavior is also easy to explain with a cognitive model. A very brief explanation goes something like this: One of our memory systems contains premotor patterns that are invoked and executed during skillful behavior. These premotor patterns are learned (stored and then modified) by repetitive conscious practicing of the skill. When these premotor patterns are consciously generated for the first time, they are stored in a memory system for later invocation. The memory traces are adaptive, and each time the skill is practiced, the stored premotor patterns are recalled and executed (described below). The conscious corrections to the induced movements that result in the more skillful behavior are used to modify the memory traces, thereby encoding the corrected premotor patterns. *Learning a skill* means the process of storing and modifying the premotor patterns necessary to perform the skill. Once the skill is learned, performing the skill is accomplished by invoking the proper sequence of stored premotor patterns. *Invoking a premotor pattern* means recalling it from memory and executing it. *Executing* a premotor pattern means the transformation of the premotor pattern into the final sequence of efferent signals necessary to perform the skill. This transformation may involve compensation for such things as speed of execution, specific forces necessary at the current time to perform the act (such as playing a piano having a heavier touch than the one used for practice), compensating for other current systematic irregularities, and so forth. One such systematic irregularity occurs when you speak with your teeth held together. The efferent patterns required for vocalization are very different than they are for ordinary vocalization. Execution of the premotor speech pattern consists in part of transforming it into the correct efferent pattern to the speech musculature to produce the desired speech sounds.

Unfortunately, insight is one process for which I have no reasonable cognitive explanation, although I do not doubt for a minute that one exists.

REFERENCES

Beurle, R. L. Properties of a mass of cells capable of regenerating pulses. *Philosophical Transactions of the Royal Society B.* 240:55–94. 1956.

van Heerden, P. J. Theory of optical information storage in solids. *Applied Optics.* 2:393–400. 1963.

Overton, D. A. State-dependent or "dissociated" learning produced with pentobarbital. In: P. Milner and S. Glickman (eds.), *Cognitive Processes and the Brain.* Princeton, N.J.: D. Van Nostrand Company, Inc., 1965.

Penfield, W., and Perot, P. The brain's record of auditory and visual experience. *Brain.* 86:596–696. 1963.

Penfield, W., and Roberts, L. *Speech and Brain-Mechanisms.* Princeton, N.J.: Princeton University Press, 1959.

Rieger, C., and Grinberg, M. *The Causal Representation and Simulation of Physical Mechanisms.* Technical Report TR-495, Department of Computer Science, The University of Maryland, College Park, 20742.

by **Eugene Charniak**

Department of Computer Science, Yale University, New Haven, Conn. 06520

How to register dissatisfaction with A.I. Ever since the beginnings of Artificial Intelligence (A.I.), there have been attempts to show that the fundamental goal of the field, the creation of an intelligent computer, is impossible. Being a specialist in A.I., I view Haugeland's paper as a fairly typical contribution to this literature.

There is one innovation, however, that I regard as positive, since it seems to be leading in the direction of a calmer and more considered analysis of the issues. Strictly speaking, Haugeland is not discussing whether or not computers can behave intelligently, but rather the tenability of a particular view of intelligence, which he calls "Cognitivism." Cognitivism, as he sees it, holds that the mind can be understood as a system (i.e., it operates through the interaction of conceptually distinct subsystems) and furthermore, that the subsystem interactions must be understandable at a suitably high level of abstraction. Thus, a computer that talks because it knows about words and their meanings would fit under this definition, but one that talked by, say, doing a quantum mechanical simulation of a person would not. Hence there is no necessary connection between the possibility of intelligent computers and the success of Cognitivism. Nevertheless, Haugeland would claim, and quite correctly, that A.I. as currently practiced is committed to Cognitivism and hence must stand or fall with it. The advantage of this shift in emphasis is that it removes the argument from the domain of A.I. (a notoriously controversial field) and emphasizes instead the strong links connecting A.I. with cognitive psychology and, although the author does not mention it, much of linguistics as well.

This discussion, as well as that dealing with the relation between Cognitivism and other kinds of explanation (in particular, reductionism), strikes me as quite reasonable. But this analysis is meant to be in the service of the second half of the paper, Haugeland's attack on Cognitivism in general and A.I. in particular, and here is where things go astray. The problem is that there are only occasional links between the two halves of the paper. Instead, we get, in the three concluding sections, three of the traditional objections to A.I.: 1) there is no proof that its assumptions are correct, 2) there exist problems with no obvious solutions, and 3) A.I. has been disappointing when compared to early predictions. In the remaining space, I will try to indicate why I (and, I suspect, my colleagues) do not find his objections persuasive.

The first of these points simply leaves me baffled. Why would anyone think that there ought to be a "correctness" proof for cognitivism, or any other paradigm (in Kuhn's sense; 1962 *op. cit.*)? Buried in any scientific discipline there are assumptions about what the world is like. In physics, for example, we assume that the world consists of particles and forces between them (or some such). These assumptions are never capable of proof and, in fact, are often found to be false. Cognitivism is neither better nor worse than any other science in this respect. A more interesting issue is the plausibility of the assumptions behind Cognitivism. I would be prepared to argue in their favor (the rest of the body is organized like a system; why not the brain?), but this would take us far beyond the current paper.

As for the second point, Haugeland sees three tough problems as casting doubt on the cognitive enterprise: moods, skills, and insights. Moods, I would argue, should be separated from A.I. and cognitive psychology, and the fact that Haugeland sees no way to do this does not bother me. I can fantasize several, but more importantly, I do not see moods as so pervasive in the tasks of importance to A.I. at this time. For example, moods have minimal influence on how I do pronoun reference, or how I distinguish my telephone from my briefcase. These are currently important problems: why things at times seem less livable is not. As for skills, the entire argument seems to rest on the observation that these are not open to conscious introspection, neither in the way they are learned (by practice, not by reading) nor in the way they are described (with difficulty), nor while being executed. But what is this supposed to prove? Nothing in the definition of "Cognitivism" says anything about the processes' being conscious, and quite rightly so. Furthermore, there is work in the A.I. literature to suggest that skills are amendable to cognitive techniques (see Stallman and Sussman on skill acquisition). Concerning "insight," it depends on what one means by the term. Sometimes Haugeland seems to mean "deep understanding," as when he comments that "many students can learn the routines for getting the right answers without the slightest insight into what's going on." We might imagine, for example, a computer that, given the input currents, can calculate what will happen in an electronic circuit, but that cannot say what any part of the circuit "is for." The trouble with this example is that there are programs that can do both. At other times, Haugeland refers to the insights behind major discoveries. While it is true that programs have not made any major discoveries, the traditional assumption in A.I. is that this is a difference of degree, not kind. (This is not obviously true, but neither is it obviously false. Furthermore, there has been some relevant work, Lenat, 1977. His program starts with very general knowledge about sets and about what makes something "interesting," and it comes up with, among other things, prime numbers. This, of course, hardly scratches the surface, but it does suggest that the problem is not unsolvable.)

As for the last criticism of Cognitivism, stagnation compared to early A.I. predictions, this argument accomplishes nothing, aside from raising the

blood pressure of A.I. types. That A.I. has not been able to solve in twenty years problems that have been around for 2,000 is not surprising. If Haugeland had claimed, instead, that A.I. had not made any significant progress in, say, the last five years, that would be relevant, but, I think, false. (In particular, his claim that there has been no progress in general problem-solving is demonstrably false.) This debate is hard enough without bringing in points that sound impressive, but are of no relevance.

One problem is that it is never clear in a debate like this what the opponents would take as a rebuttal of their positions. From my own point of view I can see that it will be very difficult, if not impossible, for Haugeland to change my mind. As he comments, to any problem he brings up. I will simply respond "Give us time." But I wish to claim that this is, in fact, responsible scientific practice. Scientists, as Kuhn has noted, never abandon a paradigm unless there is another waiting to take its place. Currently there is no plausible replacement for Cognitivism, so neither I nor my colleagues will give it up.

But this is not to say that the debate need be pointless. Paradigms do change and they change because one or a few scientists are dissatisfied with the current one. What Haugeland must do is provide enough evidence to convince a few that a shift is needed. But this evidence must be much more specific than vague comments about the problem of emotion, or the lack of progress in A.I. Haugeland must, in effect, become a specialist in A.I. He must be able to sit down with researchers and go over computer listings in order to find out why a program does as well as it does, or why not, and then using such specific, detailed information, wind a web of dissatisfaction that will eventually lead to a new Copernicus turning our intellectual world on its head. Quite frankly, I doubt that this will occur, because I really do believe in Cognitivism. But if this debate is to go further, it must be at this detailed level. It is simply the nature of science.

REFERENCES

Lenat, D. (1977) AM: an artificial intelligence approach to discovery in mathematics as heuristic search. Memo SAIL AIM–286, Stanford University.

Stallman, R. M. and Sussman, G. J., Forward reasoning and dependency-directed backtracking in a system for computer-aided circuit analysis. *Artificial Intelligence*, Vol. 9, No. 2, October 1977, pp. 135–197

by **Robert Cummins**

Department of Philosophy, University of Wisconsin, Milwaukee, Wisc. 53223

Systems and cognitive capacities. Haugeland tells us that cognitive psychology stands on two cornerstones: that psychological explanation should be systematic (in Haugeland's special sense), and that psychological states/processes should be characterized (via intentional interpretation) in terms of their symbolic content (Heston, 1977). This, I think, is quite right, though I would restrict these observations to the explanation of cognitive capacities: cognitive psychology need not be represented as claiming the whole field (see Fodor, 1975 last chapter). In what follows, I will concentrate on a certain claim Haugeland makes about the explanation of such capacities. I hope to show that the notion of a cognitive capacity and the correlative notion of an IBB (something that has a cognitive capacity) is as yet too fuzzy to bear the weight Haugeland sometimes put on it.

Haugeland criticizes the thesis that the only explanation of something's having a cognitive capacity (being an IBB) is that it *cogitates* (as I will say) – that is, it produces its behavior by, in part, "figuring out" how to satisfy the cogency conditions definitive of the capacity in question. Putting aside the possibility of inexplicable cognitive capacities, the thesis Haugeland attacks comes to this: Nothing that does not cogitate can have a cognitive capacity. Haugeland's argument goes like this: Something might have a cognitive capacity – be an IBB – yet not have "distinct functional components whose relevant interactions are confined to intentionally interpreted articulated typologies." *Such a device would not cogitate, that is, would not produce outputs by working out a rationale for them.* Thus, we cannot move from the fact that the input-output conditions of an IBB are given by specifying the conditions under which each (interpreted) output would be cogent to the conclusion that an IBB produces its output on any given occasion by working out which output would be cogent on that occasion. In short, something can have a cognitive capacity and yet not cogitate.

I have italicized a description of the move on which I want to concentrate. How is this supposed to follow from what precedes it? Haugeland plainly regards this question as trivial. Having envisioned a hologram-based pattern-recognizer, he remarks:

"What matters to us is that a pattern recognizer based on these principles would not (or need not) be an IPS. (1) There are no distinct functional components whose relevant interactions are confined to intentionally interpreted articulated typologies. That is, (2) there is nothing going on here which can be regarded as "working out a rationale" with quasi-linguistic representations. Section 6; the numbers in parentheses are mine.)"

From this, it is obvious that Haugeland regards (2) as a gloss of (1). This, I think, is a mistake: An IBB might satisfy (1) and fail to satisfy (2), that is, it might cogitate even though it has no IBB components at all. I shall give this claim a name – the Weak Thesis – because I want to distinguish it from a Strong Thesis that I suspect is true, but will not argue for, that is, that every IBB cogitates under some interpretation or other. The Weak Thesis is certainly true if the Strong Thesis is true, for there are certainly IBBs that have no IBB components, for example, a simple diode AND-gate. But investigation of the Strong Thesis would be a major undertaking, so I will restrict my attention here to the Weak Thesis.

If the Weak Thesis is true, then Haugeland's notion of an IPS is ambiguous. Something that cogitates, but satisfies (1) is, in an obvious intuitive sense, an information processor, but not a *system* (of IBBs). What we should like to mean by an IPS, I think, is an IBB that cogitates *because* it is a system of IBBs, that is, an information processor that performs its task by executing a rationalizing program whose subroutines are executed by component IBBs. The Weak Thesis can be established by exhibiting a "simple" IBB – that's an IBB having no IBB components – that cogitates (executes a rationalizing program). The fanciful mechanical devices typically described in the course of introducing the concept of a Turing Machine seem fairly clear examples (See Davis, 1958, chapter 1). Consider the Turing Machine Z defined by Table 1. Z computes x + y as follows:

Input: write x+1 ones followed by a blank ("B") followed by y+1 ones on Z's tape.
Output: total number of ones on Z's tape at stop.
Start: set Z in s–1 reading the left-most one.

Anything realizing Z will execute the program given by the flow chart in Figure 1. It seems evident that executing this program counts as constructing an output satisfying the cogency conditions for addition, and therefore cogitates. Any IBB that executes the program cogitates and is therefore an information processor in my sense. But a realization of Z need not be an IPS: it could be a simple IBB in just the way a typical AND-gate is. All this requires is that there be no proper components whose special business is to execute particular instructions in the program. And all *that* requires is that the states specified in the transition table should be states of the whole device (i.e., there is no saying which components of the device mediate which transitions). As an example, consider a beaker containing two immiscible chemicals, A and B, B being much heavier than A. Ones are represented by type-x tablets, blanks by type-y tablets. Tablets are insoluble in layer B, into which they sink if not dissolved in layer A. At the start, layer A is capable of dissolving one type-x tablet. Further x-tablets sink into layer B. A y-tablet restores layer A to its original condition. We provide input by linking up tablets in a track and tipping them one at a time into the beaker. The output is given by the number of tablets left in the bottom of the beaker.

It seems pretty clear that this device is an IBB (an adder) that cogitates (executes chart I) but has no IBB components. To deny this, one would have to argue that (i) the device has IBB components, or that (ii) the device does not cogitate. To press (i) is to risk trivializing the notion of an IBB. I cannot

Table 1 (Cummins). ***Transition table for a Turing Machine that computes x + y.***

Z	1	B
S-1	B, S-1	R, S-2
S-2	R, S-2	R, S-3
S-3	B, S-3	R, STOP

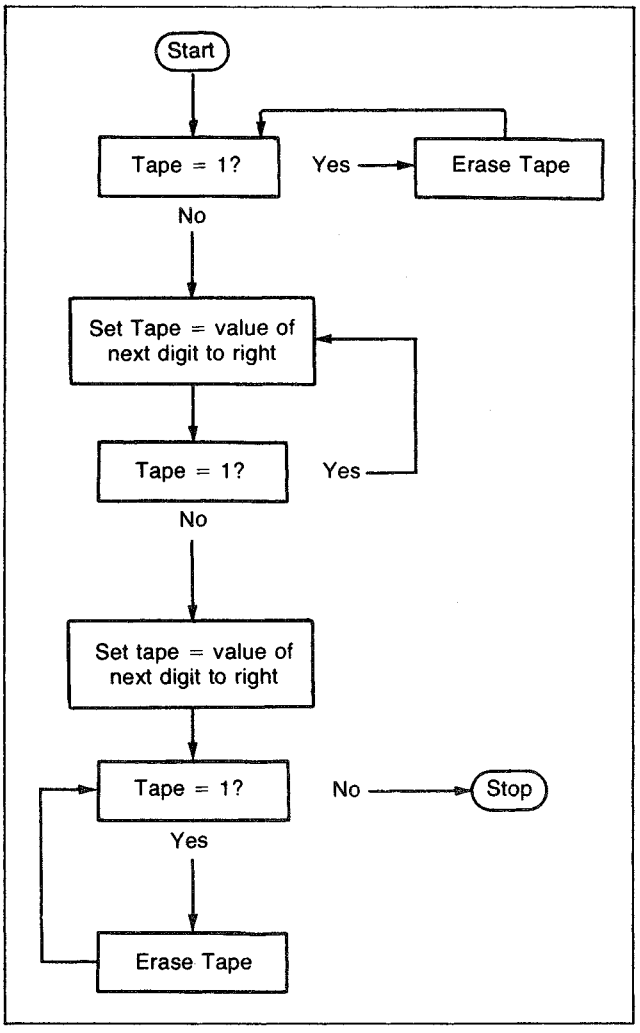

Figure 1 (Cummins). Flow chart executed by any device realizing Z.

prove this – perhaps layer-B is a "memory," though there is no retrieval – by I do not think I need to. Turing machines evidently need not be systems of IBBs any more than an AND-gate. (ii) is rather more subtle. There will doubtless be some who will hold that the beaker-device does not execute chart I, or at least not in the "right" sense, that is, a sense in which executing chart I would *"really"* be "working out a rationale" or "figuring out the answers." I think that the sense in which the beaker-device executes chart I is the only one in which program execution makes clear sense in this context: given the required interpretations, it is nomologically necessary that the device do what the chart says to do. Having already aired my views on this (1977), I will not pursue the matter here. If these points seem controversial, it is because the concepts of a cognitive capacity and of "working out a rationale" (cogitating) that are central to the cognitivist strategy are still very fuzzy at their edges – too fuzzy, I think, to bear the weight Haugeland put on them in section 6 of his paper.

A final word about the significance of all this. If the Weak Thesis is true, then "cognitivism" should not be thought committed to the view that the mind is a system of IBBs (an IPS). Cognitive psychologists attempt to explain an organism's cognitive capacities by appeal to the hypothesis that nervous systems execute certain programs, that is, by appeal to the hypothesis that they cogitate. This in itself does not imply that nervous systems are interpretable as systems of IBBs.

This is as it should be: cognitive theories no doubt give the nervous system lots of distinct cognitive jobs to do (assume lots of distinct cognitive capacities), but surely such theories need say nothing about whether or how these jobs can be delegated to physically distinguishable components. If the Strong Thesis is true, then the serious question cannot be *whether* organisms cogitate (whether they execute programs) but *how* (which programs they execute). Haugeland seems to think this trivializes cognitive psychology. That seems to me like saying we trivialize mechanics when we give up wondering whether motion is law-governed and concentrate instead on figuring out what the laws are. To suppose that organisms exercise their cognitive capacities by cogitating is not to speculate on the etiology of behavior, but to assume that its cogency is not ultimately mysterious.

REFERENCES

Cummins, R. Programs in the Explanation of Behavior. *Philosophy of Science.* 44.2:269–87. 1977.

Davis, M. *Computability and Unsolvability.* New York: McGraw-Hill, 1958.

Fodor, J. *The Language of Thought.* New York: Thomas Y. Crowell, 1975.

Heston, J. Cognitive Theory and the Content of Psychological States. Ph.D. Thesis, The Johns Hopkins University, Baltimore, Md., 1977.

by **Daniel C. Dennett**

Department of Philosophy, Tufts University, Medford, Mass. 02155

Co-opting holograms. This paper usefully dispels some of the aura of inevitability that surrounds much of the enthusiasm for cognitive science today. Even if, as is often said, cognitive science is the only game in town or the only well developed research program in psychology with a prayer of success, it *might*, for the sorts of reasons Haugeland cites, still turn out to have been a perniciously seductive dead end. Haugeland says, correctly, that the best antidote to the "what else could it be?" defense is a sketch of a "conceivably viable alternative." I wish Haugeland had been able to come up with something more conceivably viable than a hologram-inspired hunch. Holograms are marvelous information-*storers* and (in a restricted sense) information-*transformers*, and perhaps could be harnessed, as Haugeland suggests, as powerful content-driven associators or address-finders, but so far as we know now, that is the extent of their powers, and what we need are ways of getting information *processed* in a much more dramatic sense; an army with high-fidelity holographic maps of the battlefield does not yet have a battle plan, and what no one (so far as I know) has yet suggested is how to get holograms to turn the afferent-efferent corner and do the work cognitive science supposes done by decision-making systems, for instance. Principles akin to those of holography might indeed provide much needed breakthroughs in what might be called the implementation of systems management, but I do not see how the supposition that structures governed by such principles could supplant *systematic* organization altogether can be anything but a vague wish.

Haugeland's perceptive discussion of the distinction between systematic and morphological explanations in fact seems to provide the ground for an argument against just such a hope. He offers DNA replication as an example of a phenomenon requiring and permitting morphological explanation, though he grants that a finer-grained account of its mechanics might be, in his terms, systematic. Look in the other direction, though, at the *coarser*-grained account of the macroscopic phenomenon to enrich DNA's replicatory prowess contributes. The whole process of sexual reproduction surely requires a systematic explanation, without the rudiments of which DNA replication would be a baffling curiosity. Why should we have such strands in us? What function are they performing? Self-replication and versatile self-maintenance (i.e., successful cognition) are the two most sophisticated processes known to us, I would guess, and now that self-replication is beginning to be understood, after centuries of bewilderment and wild speculation, it turns out to require a systematic explanation. Haugeland does not have much to say about what *sorts* of phenomena are, in principle, amenable to the various sorts of explanation he describes. It is still compelling, in spite of his caveats, to suppose that no phenomenon as sophisticated as cognition could fail to have a systematic explanation. It might well turn out that the mysteries of *sub*processes of cognition that are currenty being attacked via the Rube Goldberg machines of cognitive science can be given nonsystematic, morphological explanations. This seems to be the Gibsonian article of faith (Gibson, 1966). Such breakthroughs would still leave the basic vision of cognitive science intact.

Haugeland's distinction between intentional black boxes (IBBs) and information processing systems (IPSs) is very important, for it permits him to raise the often overlooked question: are all IBBs also IPSs? In a footnote he criticizes an early paper of mine (1971 *op. cit.*) for "entirely leaving out the

crucial notion of an IPS." This was deliberate, and had I hit upon the term I would have been happy to have called my "intentional systems" "intentional black boxes," for that is what I meant to stress about intentional explanation. The fact that an entity is predictable via an IBB strategy shows us *nothing* directly about its internal structure or processing. My point there was, and still is, to isolate the manifest value of everyday, ordinary-folks-style intentional explanation (e.g., "He threw her the rope because he *thought* she was drowning and *wanted* to save her") from the fortunes, good or bad, of cognitive, *sub*personal theorizing (Dennett, 1969, 1977, forthcoming *op. cit.*). This lack of structural or systematic implications is just what makes personal level intentional explanation vacuous as psychology, a point with which, I think, Haugeland concurs.

REFERENCES

Dennett, D. *Content and Consciousness.* London: Routledge & Kegan Paul, 1969.

Critical notice of Fodor, *The Language of Thought, Mind.* 263-80. 1977.

Toward a cognitive theory of consciousness, In: C. Wage Savage (ed.), *Minnesota Studies in the Philosophy of Science,* Vol. IX, forthcoming.

Gibson, J. J. *The Senses Considered as Perceptual Systems.* Boston: Houghton Mifflin, 1966.

by **Zoltan Domotor**

Department of Philosophy, University of Pennsylvania, Philadelphia, Pa. 19014.

Cognitive problems and problems of cognitivism. The single shortest phenomenologist argument against cognitivism goes like this (cf. Dodwell, 1971, p. 371): I meet someone in the street. As he walks toward me, I suddenly recognize him: "It's Steve!" The Steve I am meeting now *is the same* Steve I met yesterday. How do I know he is *the same*? A quick cognitive style answer would be that my eyes register Steve's image and some physiological signals transmit it to the brain, where in turn by another physiological process it is established that the input image *fits* the stored image. Now we ask, "*How* does the process know that the input is an example of *fitting*?" What is the criterion for this sort of decision? Well, there is a second-order mechanism that tells what it is like for something to fit a stored image. In other words, one postulates a model for fitting. But then, surely, a third-order device will be required to guide the use of the model of fitting, and so on. An infinite regress has been generated in which a mechanism is "explained" by its higher level successor mechanism, but in fact nothing has been explained. Now a compressed phenomenologist answer is that when I see Steve, I do not really see Steve, but rather what I see is the-same-Steve-as-yesterday (and this I see in the same way I see color). Thus the same-as-yesterdayness is built into the perceptual process from the outset as a kind of primitive. This is how the explanation of *how* one knows sameness comes to an end, without employing an infinite series of models and decision guides.

There are three queries I have about Haugeland's most stimulating discussion of the cognitivist-phenomenologist quarrel. The first concerns the modes of explanation used in physics versus cognitive psychology. I contend that the division and comparison Haugeland makes is there only if we rigidly identify physical theories with sets of (differential) equations and cognitive theories with some sorts of networks of (information processing) functional elements (black boxes). The comparison suffers from a category mistake. What one should compare, I maintain, is on the one hand the physicist's equations with those describing the information processing network (input-output equations for functional elements plus constitutive equations relating the elements in the network), and, on the other, the physicists's system (in general, a network of interacting components) with that of a cognitivist (also a network). The point is that nomological explanation belongs to a descriptive (syntactic) level, while Haugeland's morphological explanation operates on a structural (model) level. Using the standard (logician's) structure-description completeness relationship, one form of explanation is convertible into the other. Another angle to the business of explanation I cannot discuss here is the use of counterfactuals (even in inductive situations).

My second query is about meaningfulness. Here again, the difference between, say, physical and cognitive styles of theorizing may only be apparent but not intrinsic.

S. S. Stevens and those around him handled empirical meaningfulness in terms of scale *invariance*. For example, statements concerning standard deviation of temperature are empirically meaningless because standard deviation is not invariant under interval (affine) transforms, defining temperature scales. Passing from here to more involved physical models (using several quantities), one observes that meaningfulness can be understood in terms of invariance groups of automorphisms of these models. In a model-theoretic framework this in turn corresponds to the problem of (first-order or higher-order) *definability* in the language of the models.

I suggest that meaningfulness as discussed by Haugeland ultimately comes down to the problem of definability. To say that in the case of a strange chess-playing black box "we must know what its inputs and outputs are, and how to interpret them," to me simply means that the inputs and outputs must be *definable* in our language of chess, which again means at the model level that certain *invariance* criteria must be satisfied.

A third query is about phenomenological hurdles. It may well be that skills such as wine tasting or opera singing (if these are really skills) are simply outside the domain of applicability of cognitive theories. Every known law in physics is known to have exceptions, and it does not make physics any less respectable. As long as the domain of applicability is pleasingly large, kinky exceptions will not make any difference in holding it.

Simple examples are baffling enough. Take the well-known story about Paul Dirac, traveling with a friend in a train in England (Rota, 1973, p. 11). As they were looking out of the train windows, Dirac's friend remarked: "Look at the beautiful white sheep!" Dirac answered, "Yes, white on one side!" Of course, we have to somehow argue that Dirac was wrong. We see the grazing sheep white on both sides! Cognitive and Artificial Intelligence specialists argue by means of *knowledge-driven* (top-down) models, philosophers in epistemology lean on either causal/counterfactual chains or data-, event-, concept-, or hypothesis-driven schemes. All in all, it amounts to a battle against gestalt.

The gestalt slogan "The whole is more than the sum of its parts" is to the point only when one means "any sum." A cognitivist rebuttal must be that there is always a right kind of subdivision of the whole into its parts whose correct sum *is* the whole, even though one may not always be able to find it computationally.

REFERENCES

Dodwell, P. C. Is a theory of conceptual development necessary? In T. Mischel (ed.) *Cognitive Development and Epistemology.* New York: Academic Press. 1971.

Rota, G.-C. The edge of objectivity. *The Technology and Culture Seminar at MIT.* Series of lectures delivered at MIT, October 1973.

by **Hubert L. Dreyfus**

Department of Philosophy; University of California, Berkeley, Calif. 94720

Cognitivism vs. Hermeneutics. In this impressive paper, Haugeland gives a clear and convincing account of what a mechanism is, and, in particular, what an information processing system (IPS) must be. He intends, thereby, to put us in a better position than ever before to question the plausibility of current claims that people are IPSs. Indeed, he succeeds so well in making precise the underlying assumptions of cognitivism that he seems to have produced, without realizing it, an argument that human beings cannot possibly be information processing systems. What his work shows is that the precondition for studying human beings as information processing machines, viz. that they be regarded as intentional black boxes (IBBs), may well be incoherent.

The possible incoherence stems from an incompatibility of the two conditions on what can count as inputs and outputs of an IBB – roughly that the inputs and outputs (I/Os) must be both computable and interpretable. I will try to show that, although these conditions are obviously compatible for a machine that plays a game like chess, they become contradictory when generalized to human beings. More specifically, I hope to show that Haugeland's two conditions on being an I/O of an IBB pick out two different sets of candidates, those that satisfy only condition one and those that satisfy only condition two, and that there is no room left in between for the sort of I/O that satisfies both conditions at once.

It might be helpful to the reader if I first sketch my overall point before turning to the details of Haugeland's definitions. The first condition for counting something as an IBB is that the I/Os have to be tokens of types belonging to articulated typologies. This says, in effect, that to be computable the I/Os

have to be univocal, context-free, elements. It is certainly plausible that human beings have such I/Os, viz. the digitalized input from the sense organs and the digitalized input to the individual muscle fibers, but it is not such I/Os that we interpret as making sense.

The second condition is that the I/Os be tokens of types that are interpretable. As Haugeland spells it out, this means that the I/Os have to make reasonable sense under an overall interpretation. Now we do, indeed, make coherent sense of what people say and do, but the level of description on which we do this, listening to their utterances, watching their expression, and following their gestures, cannot be further analyzed into articulated types composed of simple types. We make sense of utterances in ordinary language, but ordinary language (let alone facial expressions and gestures) is not a calculus.

One might think that the interpretable types required by condition two are simply the computable types picked out by condition one under a different description, but, although these two sets of I/Os are, indeed, token-identical in any particular case, I take it that Haugeland accepts a Davidsonian (1970 *op. cit.*) argument that due to the context-free character of the computable types and the holistic character of the interpretable types, they cannot be consistently type-identical. The incompatibility of the types defined by the two conditions is usually covered up by using such terms as "code," "stimulus information," and "precise I/O function," which seem to satisfy both conditions. It is the great service of Haugeland's rigorous account to bring out the question begging nature of all such terms.

I will now formulate this issue as precisely as I can, using Haugeland's definitions and restricting myself to the special case of linguistic behavior.

To begin with, in definition 1 (Section 3), Haugeland defines a set of types that are "*uniquely determinable* relative to a specified range of phenomena" such that "no phenomenon is ever a token of more than one types." This is to say that the types must be unambiguous, and, of course, chess notation satisfies this condition, but natural language is ambiguous and the ambiguity may well be necessary.

Natural language also fails to satisfy definition 2 (Section 3) whereby an articulated typology is defined as "an ordered pair of uniquely determinable sets of types such that: i. tokens of types in the second set of (complete types) are composed of one or more tokens of types in the first set (simple ones)." The obvious candidates for the simple types are words and for the complete types, sentences. (If we think of the simple types as letters and the complete types as words, we will run into the problem that letters are not given an intentional interpretation.) But if words and sentences *are* the candidates for the simple and complete types, they fail to satisfy condition ii., viz. "no token of a simple type ever actually occurs . . . except as a component of a complete type." Unless the inclusion of the phrase "one or more" in definition 2.i. makes that condition tautologically true, so that every use of words is a sentence; we can easily think of uses of natural language in which single words and phrases occur that are not sentences (complete types) and yet the context fills in the sense. (A famous example is the word "slab," uttered by workers in one of Wittgenstein's language games.)

According to definition 3, in an intentional interpretation of an articulated typology "the determination of what any token of a complete type means must be made *entirely* in terms of how it is composed of tokens of simple types." (My italics) But, again, this is not true of natural language, where context obviously determines speakers' meaning and, it has been argued, even determines sentence meaning (Searle, 1978).

There is no place in a typology of words and sentences for the elements of a formal theory of context, even if, as seems highly unlikely, one could ever be worked out.

All these definitions prepare Haugeland's final formulation of the computability condition, "a quasi-linguistic representation is a token of a complete type from an intentionally interpreted articulated typology." But in the light of what we have just seen concerning natural language, it is no longer clear why the representations in question are called quasi-*linguistic*, since they lack two ubiquitous and perhaps essential features of natural language: ambiguity and context dependence. If Haugeland's definitions are taken as defining "quasi-linguistic," they lead to the awkward result that natural language itself is only quasi-quasi-linguistic.

Finally, if we remove the restriction of our discussion to the consideration of language and consider other inputs such as gestures and facial expressions, these turn out to be even worse candidates for the status of quasi-linguistic representations than do expressions in natural language, since as far as we know such expressive behavior has no formal syntax or semantics, and the determination of its meaning depends entirely on what is perceived as the current situation. Indeed, it is not clear in this case what the simple and complete types could be.

It might seem that the way out of this difficulty is to hypothesize a highly abstract formal "language of thought" in which language, gesture, and the whole pragmatic context could be formally expressed. This, I think is Fodor's (1975) move, and it does satisfy the computability condition that the inputs and outputs be specifiable in an articulated typology. But, like neuron firings, such abstract structures of the digitalized input and output, even if they do exist, are not tokens of the types we interpret when we make sense of people's behavior. That is, they fail to satisfy the interpretability condition as specified in definition 5, iii, that the "outputs make reasonable sense in the context of actual prior inputs and other actual outputs." Either we must add the additional farfetched, and perhaps incoherent, assumption that we can learn the abstract language of thought so we can make sense of it as part of the everyday practices and behavior that provide our paradigm of making sense, or Haugeland's definitions have shown that, in the last analysis, the project of cognitivism is an incoherent dream dictated by the demands of giving a scientific account of meaningful behavior. Whether this conclusion reflects adversely on Haugeland's definitions or on cognitivism, I am not at this moment able to say.

What is clear to me, however, is that Haugeland should not have assumed so easily that his example of a chess-playing machine being an IBB could be "generalized," since what is at stake is precisely the difference between human context dependent behavior and a context-free game. It follows that he should not have allowed that by applying his definitions, "flexibly" normal people are IBBs, unless "flexibly" covers all the ways we have just listed in which his definitions fail to fit human behavior. If he retracts these concessions, his paper poses an even greater challenge to cognitivism than he explicitly claims. Defenders of cognitivism must either weaken Haugeland's definition of an IBB or else swallow the counter-intuitive implications of Fodor's line. Otherwise, we must remain within our common-sense intentional circle and allow that to account for action and perception we "enlarge our interpretation, and say that [people] perceive and act under descriptions (sees that . . . , intends . . . ., etc.) and regard the descriptions as inputs and outputs." In that case, since computation of these I/Os is impossible, we will find ourselves doing hermeneutics rather than cognitive psychology.

REFERENCES

Fodor, J. A. *The Language of Thought*. New York: Thomas Crowell Co., 1975.

Searle, J., "Literal Meaning," *Erkenntnis*. 1978 (in press).

by **Judith Economos**

2 Edgemont Road, Scarsdale, N.Y. 10583

Mind that last step; I think it's loose. Since Haugeland treats a subject of considerable importance, it is unfortunate that he chooses to present Cognitivism as a program, like Freudianism or Behaviorism or the other "isms" that sweep psychology from time to time. It is more useful to look at Cognitivism as an articulation and defense of a special kind of higher-level explanation.

Explanations come in kinds for one of two reasons: (1) Heuristic. Some strategies of explanation simply provide more intuitive, intelligible answers to a given sort of question than do others. (2) Ontological. Some explananda are answerable to regularities of behavior fundamentally and irreducibly different from those of other explanada. A relatively uncontroversial example of case 2 is the mathematical-logical versus the material-empirical. Causal formats are not suitable for the demonstration of mathematical propositions; and at least since Bacon, a priori explanations have not been thought appropriate for observable phenomena.

For reasons of parsimony, if not aesthetics, and unless obliged to think otherwise, we suppose these two kinds of explanation to be the only two irreducible and indispensable ones. Therefore the divisions of nonmathematical explanations into causal, functional, teleological, cognitive, and so on, are divisions for heuristic convenience only. For example, when teleological talk (about purposes, aims, goals, instrumentalities, etc.) occurs as a shamefaced shorthand in evolutionary explanations, it is considered as merely a promissory note redeemable (in principle) in impeccably "blind"

causal/statistical terms. The "correct" kind of explanation would be tediously long, repititive, circumlocutory, and unintuitive; nonetheless, it is considered the true coin.

If this same redeemability were not true for other high-level frames of explanation, like systematic and cognitive, it would imply an ontological apartheid of their explanada, hence an emergent dualism at a very high level of explanation – a disastrous result for any hoped-for Unity of Science. Haugeland concedes that "a complete reduction of psychology is one of the dreams of unified science" but is explicitly uninterested in discussing it (see, e.g., the end of his sec. 5). He does not even think that, in the light of Cognitivism, it is much of a problem: ". . . the fact that [states with significance] have some characteristics or other, independent of the interpretation (that is, they are [sic] causal interactions with the environment), means that there is no mystery about how states with significance 'connect' with the rest of nature." (end, sec. 3) But this is false. There is a mystery, if the significance of these states (and not just the states minus their significance) has causal consequences – and it must if it is not an epiphenomenon and in violation of the laws of thermodynamics. How it can have causal consequences is precisely the question. Rigorous and nonmystical explanation of this "mystery" is well worth the attention of interested psychologists. It would be good to see coherent, sophisticated attempts to elaborate logical structures capable of explaining intelligent behavior as a natural, nonemergent phenomenon. For in explanations of this sort so far there really is a missing logical link between the last Intentional step and the first non-Intentional one.

It is true that Haugeland tries very hard to bridge this gap. But does the concept of physical instantiation of an Intentional schema accomplish the bridging? Not quite yet, I think. There are several difficulties with the definition of "Intentional interpretation of an articulated typology," but their descriptions would require more space than a Commentary affords. Therefore I offer one blanket objection: Haugeland, it seems to me, tries to make the type-token relation do for the besought phenomenon-meaning relation – and the type-token relation is too fragile to bear the load. Indeed, a type – e.g., *the letter A* – is ordinarily taken to be merely the set of all its tokens. I doubt that so purely extensional a relationship can serve the desired end, which explicitly includes "making reasonable sense." Therefore (if Haugeland's definition will not do) the intuitively satisfying, the promising, the pregnant notions of interpretation and instantiation require lots more work. Can the demonstration of an isomorphism between an (abstract) intentional schema and the (abstract) type of (concrete) physical structures serve as a rule of inference? Used in a well-behaved formal scheme of argumentation, would it close the logical gap between the Intentional and the non-Intentional steps? And what would it commit us to, ontologically?

Haugeland made a brave beginning on considering these questions. I am aware that I am in part complaining that he did not write the paper I should have liked him to write. Moreover, I have ignored parts of the paper he did write. For example, I do not think it is yet time to worry about whether Cognitivism, as an imperial program, can conquer moods and skills. I am much more concerned with whether it can coherently handle anything at all. I hope it can; and I hope that for his next paper Haugeland bends his talents– and his Cognitive phenomena – to a meticulous demonstration that it can. Finally, I hope he is permanently disabused of the näive belief that in being "a science of a distinctive form," Cognitivism "sidesteps many 'philosophical' objections." There is no escape from philosophical objections but to recognize them, meet them fairly, and overcome them – to everyone's profit.

by **James G. Greeno**

Department of Psychology, University of Pittsburgh, Pittsburgh, Pa. 15260

Systems and explanations. Haugeland has made a significant contribution by beginning to characterize explanations that show how something works, rather than by showing why something happens. Haugeland asserts that these systematic explanations are characteristic of cognitive psychology, and that by explicating them he legitimates the scientific study of cognition.

I agree that systematic explanations are different from nomological explanations. However, I think they are not unique to cognitive science. I will try to argue in this commentary that the kinds of systematic explanation that dominate recent discussion in cognitive science also play an important role in traditional physical science. I am not sure whether my argument, if accepted, would make cognitive science more or less legitimate than it is on Haugeland's analysis. On the one hand, my analysis increases the rapport between cognitive and physical science because it shows that there is more in common between them than Haugeland's analysis suggests. On the other hand, systematic explanations are considerably weaker than nomological explanations, so we arrive at the familiar opinion that cognitive science is in a primitive state.

First, let me dispose of a distinction in Haugeland's analysis that does more harm than good. Haugeland asserts that there are three kinds of explanation; I think there are really only two. Morphological and systematic explanations are not different kinds of explanations. At most, morphological and systematic explanations can be distinguished on the basis of the systems that are explained. In some systems, such as fibre optics bundles, the interactions between components are simple. Then an explanation of how the system works can be given by saying how each component works, with no complicated additional analysis of how the components interact. In other systems, such as automobile engines, the interactions between components are nontrivial. Then an understanding of the system is not achieved merely by knowing how the various components function; it is necessary, in addition, to know how they affect each other, and how these interactions go together to produce the global functioning of the system. I do not think that the existence of differences in the complexity of interaction implies different kinds of explanation. Simpler systems require simpler explanations, to be sure, but I prefer to think of morphological explanations as systematic explanations of simple systems, rather than as a separate category of explanations.

Haugeland's belief that morphological and systematic explanations are different in kind has pernicious consequences. For example, it leads him to conclude that if human memory functions like a hologram, it is not an information processing system. I was shocked when I found Haugeland saying this, and I had to conclude that he had been carried away with his argument. A hologram clearly processes information. When a hologram is suitably illuminated, the illuminating energy is transformed into an image of the object that was recorded initially, and this surely must be a case of information processing. Since Haugeland cannot identify distinct components that interact to produce this activity, he cannot fit holograms into his category of things explained systematically; hence, he excludes them from systems that process information. I would prefer to have a broader concept of information processing systems.

However, to argue against the distinction between morphological and systematic explanations is to quibble. If that distinction is removed, the distinction between systematic/morphological and nomological explanation remains, and that is the central contribution of Haugeland's article. It would be hard to overemphasize the importance of distinguishing between morphological and systematic/morphological (abbreviated as S/M) explanations. Furthermore, Haugeland's reading of contemporary cognitive science is correct; S/M explanations are predominant there.

Haugeland's remarks raise the question whether S/M explanations are unique to cognitive science. His major illustration of a systematic explanation is a theory about automobile engines. I think our understanding of this problem is informed by considering physical examples that are not engineered as obviously as this one.

Consider what makes a sound. There is an initiating object that vibrates. That generates a compression wave in a medium, such as air or water. The wave is transmitted through the medium by successive compressions and rarefactions of the medium, which we can understand as increases and decreases in the density of molecules of the medium. Eventually, the wave produces a vibration of someone's eardrum, and that causes neural impulses that correspond to hearing a sound. There are problems at the point where we say that the sound is heard, but excluding that, the description seems a straightforward illustration of S/M explanation. It has components that interact in nontrivial ways, and the explanation consists of identifying the components and showing how their interaction produces the phenomena that are interesting.

Consider another example: a teakettle of boiling water on a gas stove. Components include the gas coming through the burner, the flame, the kettle, and the water. Understanding the system requires knowing how these components interact. There are quantitative properties, and deductive arguments can be given about specific features, such as the temperature of the boiling water (see Hempel & Oppenheim, 1948, *op cit.*). However, an analysis of the system into components and relationships is essential to understanding the system.

I believe that it is an error to omit the kind of knowledge involved in Haugeland's S/M explanations from a characterization of knowledge in physical science. Recent analysis of problem solving in physics by de Kleer (1975), Larkin (1977), Novak (1976), Reif (1977), and Simon and Simon (in press) have all emphasized the importance of qualitative knowledge in the process of representing and solving problems. There are strong similarities between the processes discovered in these studies and the kinds of qualitative analyses that Haugeland discusses by way of characterizing S/M explanations. The idea is developing that knowledge of physical science has two aspects: knowledge of the syntactic system in which formal deductions are constructed, and knowledge of the semantic models of the formal system. Informal, semantic knowledge of the kind represented in S/M explanations remains an important component of physical science; in fact, it appears to play a greater role in the problem-solving performance of experts than of beginning students (Larkin, 1977; Simon & Simon, in press).

Rather than considering S/M explanations as a distinctive feature of cognitive science, I think they are an important feature of all scientific knowledge, although their role in physical science may have been neglected because of the strong development there of formal knowledge and of nomological explanations. Qualitative hypotheses are empirically testable, but to the degree that knowledge in a science is nomological, hypotheses are more precise and more strongly testable. For example, a qualitative model of sound is sufficient to explain why sound will not be propagated through a vacuum, and that model would be refuted if sound did travel through a vacuum. However, a more precise and formal model can be stated, including the speed of propagation through various substances, and this permits stronger tests. Haugeland makes a correct and important point in noting that S/M explanations are based on legitimate, empirical knowledge. On the other hand, cognitive science, in which almost all of our present knowledge is of the qualitative kind, is less testable than physical science. Whether more formal, nomological knowledge will eventually develop in cognitive science remains to be seen. I do not believe that the nature of our subject matter limits us to qualitative knowledge and S/M explanations in principle.

REFERENCES

de Kleer, J. *Qualitative and Quantitative Knowledge in Classical Mechanics.* Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Technical Report AI-TR-352, December 1975.

Larkin, J. H. *Skilled Problem Solving in Physics: A Hierarchical Planning Model.* Unpublished manuscript, University of California at Berkeley, September 1977.

Novak, G. S. Computer understanding of physics problems stated in natural language. *American Journal of Computational Linguistics.* Microfiche 53, 1976.

Reif, F. *Problem Solving in Physics or Engineering: Human Information Processing and Some Teaching Suggestions.* Berkeley, California: Group in Science and Mathematics Education, University of California, October 1977.

Simon, D. P., and Simon, H. A. Individual differences in solving physics problems. In: R. Siegler (ed.), *Cognitive Development: What Develops?* Hillsdale, N.J.: Erlbaum Associates, in press.

by **R. Harré**

Linacre College, Oxford University, Oxford, England

Half-way to realism: Some sympathetic comments on Haugeland's defence of cognitivism. Haugeland's interesting paper raises, for me, two important and interlocking issues, one in the philosophy of science and the other in the methodology of psychology.

1. Given the shortcomings of logical positivism (empiricism) as a philosophy of science, what account of science should replace it? I shall argue that though Haugeland's philosophy of science is a step in the right direction, it is only a partial revision of logical empiricism and preserves at least one of its undesirable features. But Haugeland's admission of "morphological" and "systematic" explanation-formats ought to have carried him all the way to the neorealist position.

2. In admitting cognitive states, contents, processes, and so on, back into psychology as real entities with causal powers, the methodology proposed in Haugeland's theory as a grounding for current practices in psychology is still scientistic,[1] since it draws on only one of the several possible overall models of human functioning that become available in abandoning logical empiricism. The effect of a scientistic choice of overall model is to narrow unnecessarily the range of rhetorics available for psychological explanations.

The philosophical issues. Two deep philosophical questions are raised by Haugeland's concessionary explanation-formats. They are: (1) Whether the explanatory power of a theory derives from its content or from its form. (2) Whether sequential regularities between events and states are the proper and only basic content for laws.

To make clear how these questions arise, I must turn for a moment to compare and contrast the logical empiricist conception of science with the neorealist theory. Logical empiricism had two components: empiricism and logicism. Cognitive psychology seems to have broken successfully from empiricism, that is, it is prepared to tolerate concepts whose empirical content is not given in immediate experience. But it still retains traces of logicism, since, so far as I can see, even the liberalized conception of scientific explanation offered in Haugeland's paper retains a basic foundation of laws of regular sequence, or something like it.

The neorealist position is a still more radical departure from logical empiricism. Neorealists (e.g., Bhaskar, 1975) hold that the explanatory power of a theory derives neither from its form, as the logical empiricists held (Hempel, 1965), nor from its ability to assuage psychological needs of one sort or another (Scriven, 1962), but from the content of the theory and what it says about the world. In particular, in order for a theory to have explanatory power it must describe plausible generative mechanisms and their modes of working such as to produce observable patterns of, say, human action and talk.

But to explain an observable pattern two questions need to be answered – their distinction is lost in Humean reductions of the causal mode of explanation to regularity of precedence. (1) What conditions bring about (that is, bring into existence) the patterned products we observe. To this question a reply must mention both an agent or agents and the releasing conditions for its activity to be realized in action, or, where the dynamic element in a causal process is external to the individuals and substances involved, some stimulus. (2) But as a scientist one must also be prepared to answer the question: "What was responsible for the properties of the patterned product generated by the released activity of the agent or by the external stimulus?"

In general, answers are to be found in structural properties of the material conditions of the production of the patterns we observe. One might call these "templates."

In some cases, template and causal agent are materially identical. It is the gravitational field that both produces downward acceleration and is so structured that the motion has a certain definite form. In other cases, template and agent are materially distinct, as when a person acts in an orderly fashion by following a publicly promulgated rule.

Haugeland's "morphological" and "systematic" explanation-formats are none other than special cases of the general neorealist schema. It is worth remarking parenthetically that the "subtle" difference he notices between them is exactly what Polanyi (1969) noticed between the two kinds of boundary conditions whose imposition on the laws of physics and chemistry yields integral systems. In a sense Haugeland has discovered for himself, by another route, the theory that the neorealists have been arguing for all along (Harré, 1970). But once that schema is admitted as legitimate, the Deductive-Nomological (D-N) schema cannot be tolerated as an adequate explanation schema at all. At best, it is reduced to a peculiar and degenerate case of the neorealist form of explanation.

There are two common philosophers' mistakes in Haugeland's account that may have concealed from him how radical is his admission of non-"D-N" explanation-formats of just the kind he has proposed. He says, "The most familiar scientific explanations come from classical mechanics." "Familiar to whom?" one might ask. I think the answer must be "To philosophers of the logicist tradition," since only in classical mechanics does one find pieces of reasoning that are much like the D-N format. Explanations in chemistry, physiology, geology, genetics, and linguistics are far more "familiar" to the scientific community at large. Even though they bear little resemblance to the D-N format, they are one and all intelligible and correctly formulated according to neorealist criteria of explanatory power since they are constructed around a certain kind of content.

The second matter that leads Haugeland astray is his rather casual use of the term "model." I fear it shows that he is not immune to familiar confusions about the way models function in theories. "Morphological explanations," says Haugeland, "are often called 'models'." On the contrary, one should say

Table 1 (Harré).

Content	Observed Pattern (OP)	Generative Mechanism (GM)	Model of GM (M)	Source of M (MS)
Epistemic Status	Known	Unknown	Hypothesized	Assumed

that such explanations describe models, and from that thought the neorealist explanation schema for the mapping of the necessary content emerges directly (see Table 1).

The structure is linked together through the conditions that (1) M must behave analogously to the way GM is known to behave, that is, hypothetically yield OP, and (2) M must have a nature analogous to MS, that is, it must hypothetically belong to the same ontic category as MS.

This schema not only correctly identifies that epistemic status of the components of the content of an explanatory theory, but allows us to formulate a reasonable account of how, through its two analogies, the model M is constructed in the absence of knowledge of the generative mechanism GM. Corresponding to the material relations of analogy among the components of content there are semantic relations of metaphor and simile among the concepts and their symbolic vehicles in the discourse in which the items in the scheme are described, that is in the explanatory theory. In this way we can account for the experience-anticipating intensional content of transexperiential concepts.

The methodological issues. In constructing a psychology, we need a source of concepts to enable us to conceive of appropriate models of the generative mechanisms we currently cannot study empirically. Cognitive psychology, as sketched by Haugeland, clearly depends upon a very general "computer-cum-machine intelligence" source for concepts and their interrelations for constructing a description of causal (hidden) processes that could be responsible for manifest human performances.

In terms of iconic model theory, as I have sketched it above in the neorealist schema, psychological hypotheses of generative processes are formulated first in terms of thinking, feeling, and so on. To link them with the physiological base, they are then transformed by being redescribed in information-processing rhetoric. This rhetoric is not derived from the descriptive language of thinking, feeling, planning, and so on, by abstraction, but from the model's source. This leads to a cybernetic representation intermediate between the cognitive representation of the processes that generate thought and action, public and private, and the physiological mechanisms that realize the cybernetic representation and the structures and processes it purports to describe. And, of course, it is easy to complain that the choice of information-processing as a model-source is gratuitous. Since the descriptive system for iconic models of generative "mechanisms" derived from that source drastically reduces the resources for conceiving of such "mechanisms," it is not only gratuitous (that is, uncalled for by the nature of the individual substances present and their behaviour, the things whose nature and behaviour we are studying), but coming from a general scientism in the intellectual environment, it is likely to be positively wrong.

By way of contrast, the ethogenic movement, now becoming the dominant force in social and personality psychology in Europe, would admit no other model source than human beings themselves (Hudson, 1975) and no other rhetoric than that derived by a critical analysis from elaborated forms of ordinary language. In practical terms, this methodological principle leads to the setting up of literary productions as sources of concepts for formulating psychological theories, since literary productions already encapsulate, *implicitly*, widely tested psychological (and microsociological) theories about the genesis and meaning of human action, public and private, cooperative and agonistic.

The most striking developments of this idea have been the dramaturgical approach in social psychology associated with the Oxford school of social psychologists, sometimes called "accounting theory" (Harré, 1977) and the autobiographical approach to personality studies originating in the Brussels school of assessment of criminal personality (De Waele, 1976). Both draw upon literary productions for their concepts, and upon elaborated and critically tested ordinary language for their rhetorics.

Accounting theory depends upon the interaction between two model-sources. Human action and interaction is analyzed within a general framework derived from the way drama productions are staged in both traditional (formal episodes) and radical (improvised episodes) theatre. This analytical and explanatory system interacts with the adoption of the widespread human practice of reworking social and psychological reality by talk to make the social world and human action in it appear reasonable and proper, and the people engaged worthy of respect, (Marsh, 1978). The theories people use to promote this picture are, on the ethogenic theory, both revealed in their talk and accounts and operative in the way they formulate their intentions and plans, and hence in much of what they do. This aspect connects with the recent attempt by Gauld and Shotter (1977) to transform the previously progammatic ideas of hermeneutic psychology into a practical technique.

In short, there are now two psychologies in existence that could reasonably claim to be alternative to the old operationalist/experimentalist approach. Both have arisen in conscious opposition to behaviourism in both its ordinary and its radical forms. Cognitive psychology has transcended the crude empiricism of its predecessor that confined the experimental tradition to the status of a protoscience or pseudoscience, most aptly compared to medieval alchemy. Cognitive psychology is a genuine science. But from the larger standpoint it is still scientistic in its conceptual resources. In a word, it has "gone cybernetic" too soon.

At this point another of Haugeland's philosophical contributions becomes relevant, namely his perception that his morphological and systematic explanation-formats explain through identifying underlying forms that are responsible for the form of the products of certain generative processes. In the morphological case, a synchronically existing form is responsible for the synchronic form of the product. In the systematic case, a synchronic form (the structure and layout of an engine) is responsible for the diachronic form of its operations and the structure of its movements in time. He is clearly right to see this as a radical shift from nomological foundations of explanation, since there are no laws of nature relating specific forms to specific forms. Yet to cite a specific form, isomorphic with the form of a product, as responsible for that form, is clearly immediately intelligible. Involved here, is what one might most properly call a "Principle of Natural Order," stating that structured products are generated from structured templates. It is a principle of the greatest importance in contemporary science that pays such great attention to structure. But the perception of the central role of such a principle in the methodology and metaphysics of science goes back to Aristotle and his idea of formal cause. Again Haugeland has independently and by another route discovered one of the central tenets of ethnogenic psychology – that structured products must have structured causes (Harré, 1977).

My last point concerns Haugeland's way of dealing with the necessity of grounding psychological functioning in physiological processes. Again I want to argue that his argument points in the right direction but could have been carried further. He correctly points out the impossibility of a theory of neuronal bases for cognitive functioning that takes as its neuronal units structures and processes identified according to physiological criteria alone (or any other arbitrary criteria). But he regards as circular the use of "reductions from above," that is, criteria of individuation for physiological entities derived from psychological criteria for individuating associated psychological processes and states, and so on. This causes him some unease. But when we realize that what Haugeland calls "reduction from above" is the familiar taxonomic priority thesis, proposed, for example, by Jensen (1972), it is easy to put his mind at rest on this score. Though preliminary classifications of physiological entities relevant to psychological functioning must be derived from psychological criteria, the relation so established is between taxonomies, that is, between kinds, not between individuals. So, though the kinds are circularly related, individuals may come on the scene that cannot be classified in either the existing psychological kinds or the existing physiological kinds. Occasions could occur in which individual states and processes of either category would threaten the wholesale application of the taxonomic priority theses (TPT). Once a physiological kind has been established, by whatever means, including TPT, it is surely logically possible that an instance of it could exist without a report of an instance of the corresponding psychological kind being given. The appearance of such a case is usually dealt with in practice by forming a disjunctive taxon, saying, for example, that anxiety is correlated either with adrenalin or with tryptamine, and postulating a hidden physiological parameter to account for the likeness or identity of the psychological reading of the state (Harré, 1970).

It is worth noticing that identifying the physiological "machinery" of psychological functioning by the use of the Taxonomic Priority Thesis, that is, by the method of identifying physiological kinds from already established psychological kinds, does not depend on any particular model-source for psychological theory. Cybernetics is attractive, as a model-source, since the brain is a structure of some sort, with electrical and chemical processes occurring in it. Such a model-source has a special priority on the grounds of plausibility, since the material realization of computing systems also occurs in a structured thing in which electrical and chemical processes occur. But it is clear that the use of TPT to ground psychological functioning in the brain and nervous system is compatible with quite other sources for models of the unknown processes that generate meaningful human action. And, as I have pointed out, the ethogenic movement, the European parallel to cognitive psychology, draws on literary sources for its models of generative processes, models that through TPT could lead to a conception of the brain as a theatre for the rehearsal and production of dramas, and many other things besides.

NOTE

1. Scientism is the assumption that a new field of enquiry becomes scientific by drawing on established sciences for its model-sources.

REFERENCES

Bhaskar, R. *A Realist Theory of Science*. York: Alma, 1975.
De Waele, J. P. In: R. Harré (ed.) *Personality*. Ch. 7. Oxford: Blackwell, 1976.
Gauld, A., and Shotter, J. *Human Action and its Psychological Investigation*. London: Routledge & Kegan Paul, 1977.
Harré, R. *The Principles of Scientific Thinking*. London: Macmillan, 1970.
The Ethogenic Approach: Theory and Practice In: L. Berkowitz (ed.), *Advances in Experiment-Social Psychology*. pp. 284–314. New York: Academic Press, 1977.
Hempel, C. G. *Aspects of Scientific Explanation*. New York: Free Press, 1965
Hudson, L. *Human Beings*. London: Jonathan Cape, 1975.
Jensen, U. J. 'Conceptual Epiphenomenalism'. *Monist*. 56:250–75. 1972.
Marsh, P. In: Marsh, P., Rosser, E., and Harré, R. (eds.), *The Rules of Disorder*. Ch. 4. London: Routledge & Kegan Paul, 1978.
Polanyi, M. *Knowing and Being*. London: Routledge & Kegan Paul, 1969.
Scriven, M. *Minnesota Studies*. vol III, 70–230. Minneapolis: University of Minnesota Press, 1962.

by **P. J. Hayes**

Department of Computer Sciences, University of Essex, Colchester, Essex CO4 35Q England

Cognitivism as a paradigm. This paper contains much of value and deserves a more extended analysis and response than there is space for here, so I shall concentrate only on some important criticisms.

Is cognitive science empirical? Haugeland claims to have shown that "meaningfulness can be dealt with empirically." I am less sanguine than he on this point.

His argument, as I understand it, is this: on coming across a device or creature that behaves in some way, one can make hypotheses about what its behaviour means. These hypotheses can be tested empirically by seeing whether the behaviour "makes reasonable sense" under the hypothesised interpretation. There are many difficulties in this scheme. There is the Quinean point about translational indeterminacy (Quine, 1960 *op. cit.*). There is the question of how one would ever come to hypothesise meaningfulness of the behaviour unless one in some sense already understood it. There is the difficulty of distinguishing between, on the one hand, a device that speaks language L, but is somehow faulty or incompetent or sick, and, on the other hand, a device that speaks some other language L'. (If a programmer exhibits a chess-playing program that turns out, when interpreted by his rules, to make illegal or crazy moves, one concludes there is a bug in it. The alternative hypothesis – that it speaks a different language – is so unlikely as to be almost inconceivable. Again, consider schizophrenic language: is this a new form of communication, as Laing believes, or a symptom of some inner derangement?) There is the difficulty of refuting hypotheses of vague communications, for example, inanimate oracles such as the *I Ching*.

The main point, however, is this: Even if Haugeland's argument were correct, all it would show would be that *ethology* could be empirical, not that cognitive science is. Cognitivists are not faced with a black behaving box requiring interpretation (like an insect whose strange dances need to be understood), but with a box whose behaviour is perfectly meaningful, the problem being to explain how it works. Our hypotheses are not of the form "this behaviour means such-and-such," but "this cognitive ability (speaking correct English, recognising faces, playing chess) is realised by the following computation: . . . ." Haugeland does not discuss the difficulties in testing *this* kind of hypothesis empirically, and I think they are very great (see my commentary on Pylyshyn, BBS I:1, 1978).

Systems, levels, and dimensions. Haugeland is correct to draw attention to what he calls change of dimension and its importance in describing complex systems – his reply to Lucas is well taken, for example (although one of Lucas's antecedents is almost certainly false, so his conclusion would be unsupported in any case). But the concept is not adequately analysed here.

(a) The examples of systems offered are all physical mechanisms of one sort or another. And his definition of IPS is also oriented that way: thus components have between them *causal* relationships that are *interpreted* in a typology. But computer programs are not physical mechanisms and are not *like* physical mechanisms. Their "parts" are directly symbolic and interact by the transmission of information rather than energy. They do not causally influence one another: their intercourse is governed by the very rules that define the meanings of the symbols they exchange. The fact that evaluation of the LISP expression (CAR '[A B C]) results (via a call of the function CAR) in the atom A follows from the rules of LISP, not from any physical law. Nor are these LISP expressions in any sense coded descriptions of some underlying physical mechanism, of whose internal workings a causal account can be given. While of course there is, eventually, a physical computer on which the program is running, the causal relationships that obtain between the physical parts of this computer and the LISP-ish relationships among the parts of the program bear no necessary, or even law-governed, relationship one to another.

(b) Haugeland argues that the components of a chessplayer, for example, must be talking to one another about chess. He concedes that the inner vocabulary might properly contain – be richer than – the outer one. But he fails to notice that this richer inner vocabulary may include concepts that he would want to attribute to a lower "dimension" (usually called "level," as in "high-level programming language"), thus making mincemeat of his level/dimension contrast and of his main point. Minimaxing in game playing is a good example. This is an algorithm, used by almost all chess-playing programs, for searching a tree of possible moves. It depends crucially on the comparing of numbers and operations on labelled trees, neither of which are about chess. These are top-level components of the main chess-playing algorithm: they are not to be explained away as lower-dimension implementations – "intentional instantiations" – of some high-level subroutine. Minimaxing is an algorithmic technique that directly relates zero-sum two-person games, trees, and numbers.

Haugeland's central conclusion, that programs *must* "work through a rationale" expressed entirely in the vocabulary of the original "problem," is just false, although some programs may do so.

Rationales, skills, and holograms. This "rationale" idea is asked to bear a lot of weight in Haugeland's later discussion. There is a subtle shift in its meaning as this discussion advances from merely (a), the belief that the components talk to one another in the language of the problem – in any case often false – to the much stronger (b), the belief that the components explicitly *justify* the "solution," which is the final output, where this seems to mean something like explaining to themselves *why* the solution *is* a solution. It is important to see that (b) is a far stronger claim than (a), and it is (b) that Haugeland finds so implausible, applied, for example, to skills and moods. But (b) is even more wrong, as a theory of how programs operate, than (a) is.

One must distinguish an algorithm that works from an explanation of *why* it works. (I am indebted to David Marr, who finally convinced me of the truth of this.) Take an example suggested by Seymour Papert: an algorithm for catching a ball. One might expect that the rationale behind this skill involved differential equations, laws of gravitation, and so forth. But the algorithm is very simple. If the image of the ball moves upwards, run backwards; if downwards, run forwards; if sideways, run in the same direction. When the subtended angle of the image visibly increases, get ready to catch the ball. Now, this algorithm works, but it works through "no rationale," it just does its own thing. To show that it works requires some further discussion, but that is not part of the algorithm, and is not being elaborated in the heads of suc-

cessful ball-catchers. (Nor does it involve differential calculus.) And, again, most algorithms are like this.

Much of Haugeland's discussion of skills and moods thus rests on sand. For example, there is no reason at all to suppose that a computational account of chess-masters' skill need involve explicating the "thousands of games" that gave rise to the skill. Berliner (1973) discusses this problem very insightfully.

There are two responses to be made to Haugeland's "holographic" counter-example and his discussion of skills.

(a) Leaving aside the fact that a pure holographic device is at best an associative memory, and therefore could not conceivably be a whole chess player, the proposed contrast between holographs and computations is faulty. A physical holograph (optical, for example) is a physical instantiation of a certain kind of computation, which can be realised on certain kinds of parallel computer architecture (such as are found in the mammalian retina and cerebellum). Computations of this general family have been put forward as models of low-level vision (Marr, 1975) and memory organisation (Fahlman, 1977), for example. So, even low-level, wholly unconscious, fast skills may plausibly be mediated by computations.

(b) On the other hand, many apparently unconscious skills have, by a process that interleaves introspection with implementation, been teased out into systems of explicit rules. One implements the expert's first faltering introspection, then has him criticise the resulting (bad) program, then implements his suggested refinements, and so on. Typically, the expert's talents at introspection become sharpened as the process is iterated, perhaps scores of times. The first program built this way, the Stanford DENDRAL, is now the world's expert at analysis of certain mass spectra. Other programs diagnose plant diseases or recommend antibacterial therapy, with human-level performance (see Feigenbaum, 1977, for a survey). But the key point, for us, is that the human experts recognise the rules as convincing accounts of their own thinking, and the transcripts of the program's activity as reasonable, convincing justifications for its conclusions.

The success of these programs, and the way they were built, argue strongly that the experts were working through a similar rule-govered rationale in deploying their skills. Haugeland doubts whether skills are mediated by unconscious systems of rules. But the evidence is overwhelming that many "intellectual" skills are, since the rules are there to be found.

Conclusion. Haugeland clearly comes to bury cognitivism, not to praise it. Many of his unfounded criticisms of A.I. progress and prospects deserve detailed replies for which there is no space here; but there is, in any case, a more fundamental point. Cognitivism is not a theory for which positive and negative evidence can or should be sought. It is a paradigm, an approach to constructing theories, a style of theorising. It cannot be correct or incorrect, only more or less productive. As Haugeland observes, it *may* ultimately turn out to be sterile. And it *may* be very productive. But the only way to tell is to try doing normal science within it, and see where we get. Anticipating possible paradigm failures before the paradigm has crystallised is not science but antiscience.

REFERENCES

Berliner, H. *Some Necessary Conditions for a Master Chess Program.* Proceedings 3rd International Joint Conference on Artificial Intelligence, Stanford, 1973.

Fahlman, S. *A System for Representing and Using Real-World Knowledge.* Ph.D. dissertation, Massachusetts Institute of Technology, Cambridge, Mass., 1977.

Feigenbaum, E. A. *Themes and Case Studies of Knowledge Engineering.* Proceedings 5th International Joint Conference on Artificial Intelligence, Massachusetts Institute of Technology, Cambridge, Mass., 1977.

Marr, D. *Analysing Natural Images.* Memo 334, A. I. Laboratory, Massachusetts Institute of Technology, Cambridge, Mass., 1975.

by **Robert J. Matthews**

Philosophy Department, Harvard University, Cambridge, Mass. 02138

Two remarks on the characterization of IBBs. Haugeland's paper is a rich source of ideas for discussion. In my brief comment I shall focus on two difficulties that I see in his characterization of Intentional Black Boxes (IBBs). I shall conclude with some remarks about his assessment of the plausibility of cognitivism.

It is a cardinal tenet of cognitivism that the behavior of a cognitive entity is *meaningful.* As Haugeland puts it, "the observed outputs consistently 'make reasonable sense' in the context of other observed inputs and outputs." Critics of cognitivism find this tenet contentious: they fail to see how meaningfulness or significance can be made empirically respectable. Haugeland attempts to answer this objection in section 3 of his paper by means of his definition of IBBs. He summarizes his response as follows: "although meaningfulness is not an intrinsic property of behavior that can be observed or measured, it is a characteristic that can be attributed in an empirically justified interpretation, if the behavior is part of an overall pattern that 'makes sense'." Yet it is unclear why Haugeland thinks that he has successfully answered the objection. After all, definition (5) leaves totally unexplicated the crucial notion of "making reasonable sense," which to my ears sounds very similar to the notion of meaningfulness with which we began. The problem, as I see it, is that Haugeland requires that the inputs and outputs of IBBs be interpretable as "quasi-linguistic representations," but he fails to provide a semantics for these representations. Or at least he fails to show that such a semantics can be provided. His definitions (1) through (4) stipulate, in effect, that quasi-linguistic representations are syntactic structures that obey a Fregean compositionality principle (viz., the meaning of any token of a complete type is a function of the meanings of the tokens of simple types of which it is composed, where the functions in question are defined over the syntactic structures of complete types). But, as Haugeland acknowledges, nothing is said in these definitions about how the meanings of simple types are assigned; hence, we are given no semantics for these representations. Haugeland takes this to be a minor problem, since in practice we can generally agree on what does and does not "make reasonable sense," that is, we can agree on what does and does not constitute a reasonable interpretation of these representations. We can, moreover, interpret these representations by translating them into some antecedently understood language, perhaps English or Mentalese.

Critics of cognitivism are likely to be unmoved by such practical considerations. Translation, they will argue, is fine for a while, but eventually the cognitivist must get down to the task of "making sense" of the antecedently understood language. Of course, it is not incumbent on Haugeland that he provide a semantics for this language; that is a task for linguists and psychologists. Yet if Haugeland is to meet his critics' objections, and this is apparently something that he wishes to do, then it is incumbent on him to provide reasons for believing that the cognitivist enterprise will not impale itself on this task. The problem, though, is this: what sort of reasons would Haugeland's critics find compelling? What would convince them that meaningfulness or significance can be dealt with empirically? Actually providing the required semantics might turn the trick, were it not that all semantic theories that have been proposed for the task are decidedly cognitivist (e.g., Anderson, 1976; Bobrow & Winograd, 1977; Charniak & Wilks, 1976; Minsky, 1975; Winograd, 1972, 1976). But given this fact, I suspect that nothing short of what Haugeland (misleadingly) calls a "complete reduction" of cognitive psychology would satisfy these critics, since nothing short of that would demonstrate the compatibility of cognitivism with token physicalism (Fodor, 1975, pp. 9–26). But this is hardly something that can be provided in a single paper, much less by a single definition.

These skeptical objections aside, Haugeland's own explication of cognitivism requires that he be able to provide a semantics for quasi-linguistic representations. The notion of a "change of dimension," characteristic of intentional instantiation, presupposes that one is able to decide whether two IBBs "pertain to the same subject matter or problem." Intentional instantiations, we will recall, involve reinterpretation, that is, change of subject matter or problem. This presupposed ability in turn presupposes that the inputs to and the outputs from IBBs are in some fairly strong sense representations *of* or *about* a particular subject matter or problem. Spelling out this sense will require a semantics for quasi-linguistic representations. If such a semantics could not be provided, Haugeland's notion of intentional instantiation would collapse.

My second criticism of Haugeland's notion of IBBs is directed toward his distinction between IBBs and Information Processing Systems (IPSs). IPSs, we are told, are IBBs whose input/output abilities can be explained in terms of the abilities of component IBBs without having to reinterpret their inputs and outputs (see his section 4). The abilities of IBBs that are not IPSs cannot be so explained. Haugeland's distinction is tenable; however, it obscures important differences among IBBs, differences that were well captured by

Dennett's (1975 *op. cit.*) "committees of homunculi" metaphor: IBBs can be approximately ordered on a continuum from those having no componential structure (e.g., the *AND*-gate) to those having a rich componential structure (e.g., the human mind). IBBs that are not IPSs are a limiting case. I speak of the "componential structure" of IPSs in order to emphasize that having component IBBs is typically part and parcel of the interpretation of an object as an IBB of a particular type. In other words, the cogency conditions for a reasonable interpretation of an object's inputs and outputs will typically mention various internal states of that object (e.g., its intentions, desires, beliefs, etc.). Of course, having a particular componential structure is not a *criterion* for interpreting an object as an IBB of a particular type, but having a particular componential structure is entailed by the interpretation.

The following is the point of emphasizing that IBBs exhibit differing amounts of componential structure and that the possession of a particular componential structure will be one of the cogency conditions for interpreting an object as an IBB of a particular type: cognitivism is not simply a commitment to the claim that "the mind is to be understood as an IPS," but rather that the mind is to be understood as *an IPS having a rich componential structure.* It is not simply that cognitivists think that the abilities of human IPSs can be explained in terms of the interaction of components "working through" a rationale for whatever outputs they collectively produce. They also believe that explanations in terms of such rationales will be sufficiently rich in detail to provide genuine illumination and understanding; IPS-explanations will provide the explanatory force in a mature psychological theory. Haugeland's ungraduated distinction between IBBs that are IPSs and those that are not obscures this belief. Haugeland thus overlooks what is surely the most serious threat to the cognitivist enterprise, namely, IPS-explanations of human abilities may prove to be so impoverished as to be uninteresting. It is precisely this threat of explanatory poverty that lends importance to current attempts to extend the domain of cognitivism beyond that explicitly licensed by the cogency conditions on IPSs: cognitivists attempt to enrich the componential structure of the IPSs that they study by postulating "unconscious" information processing. If such structurally enriched IPSs prove unable to model satisfactorily the behaviors that they are intended to explain, then cognitivism will have shown itself to be explanatorily impoverished. The explanatory burden of a mature psychological theory will then have to be shouldered not by the IPS-explanations of that theory but by its instantiating explanations.

Haugeland's discussion in section 7 of skills and understanding is intended to suggest that attempts to extend the domain of cognitivism in this way will fail. Perhaps Haugeland is right; it is too soon to tell. But even if he is, that will not show (as Haugeland seemingly believes) that cognitivism is false. For given the firmly entrenched place of cognitivism in common-sense psychology, we would surely reconstrue cognitivist explanations in such a way as to make them "approximately true" under succeeding theories. Of course, the eventuality that Haugeland envisions would show cognitivism to be trivial, and that would hardly be less momentous than showing it to be false.

REFERENCES

Anderson, J. R. *Language, Memory, and Thought.* Hillsdale, N.J.: Erlbaum Associates, 1976.

Bobrow, D. G., and Winograd, T. An Overview of KRL, a knowledge representation language. *Cognitive Science.* 1:1:3–46. 1977.

Charniak, E., and Wilks, Y. *Computational Semantics.* Amsterdam, North Holland, 1976.

Fodor, J. A. *The Language of Thought.* New York: Thomas Y. Crowell, 1975.

Minsky, M. A framework for representing knowledge. In: P. Winston, (ed.), *The Psychology of Computer Vision.* New York: McGraw-Hill, 1975.

Winograd, T. Towards a Procedural Understanding of Semantics. *Revue Internationale de Philosophie.* 117–118:260–303. 1976.

by **Grover Maxwell**

Minnesota Center for Philosophy of Science, University of Minnesota, Minneapolis, Minn. 55455

Cognitivism, psychology, and physics. After reading Haugeland's excellent and stimulating paper several times, I found myself wondering whether there was enough disagreement between us to provide the basis for a significant critical commentary. I agree with him that cognitivism is an important and, *perhaps,* quite a promising approach to certain areas of psychological inquiry. However, I italicized the "perhaps" because I share all of Haugeland's doubts and misgivings about its probable success and, especially, about its scope. Haugeland appreciates and emphasizes ". . . the depth and scope of psychology's pretheoretic purview," and I share his intimated doubt that cognitivism can successfully cope with problems outside a rather narrowly limited portion of this purview.

One of the two main tasks of this commentary will be to argue that the limitations of cognitivism seem to be even more severe than Haugeland indicates. The other task will be to explore what Haugeland calls "modern wisdom," to the effect that psychology is, in principle, completely reducible to physics. The discussion will, I hope, reveal that this latter task is not entirely irrelevant insofar as the limitations of cognitivism are concerned.

Regarding the first point, consider what Haugeland calls the ". . . long and tortured tradition in philosophy for distinguishing two kinds of mental phenomena: roughly, cognitive or intellectual states versus felt qualities or the purely sensuous given." In other terms, this is the distinction between "sapience" and "sentience" or, in still others, between perception, cognition, and ratiocination (all cognitive) on the one hand, and sensations, "raw feels," and so forth, on the other (Feigl, 1961, 1967). This distinction may suffer from borderline fuzziness; in fact, it may well be that every (conscious) instance of sentience is tainted with some degree of sapience and vice versa – or, even, that sentience (sensation, etc.) is one of the principal vehicles for sapience (cognition, etc.). In spite of this, however, it seems undeniable that the distinction is not only viable but also that it is of fundamental importance. Now, information processing systems, as such, and, therefore, cognitivism, are concerned only with the sapient (cognitive) aspects of mental processes. It is true, as Haugeland suggests, that ". . . noncognitive [sentient] states . . . [may] . . . be effective . . . [cognitively] . . . insofar as they . . . generate quasi-linguistic representations ('red there now,' 'left foot hurts') which can be accepted as inputs by the cognitive IPS." But, in any given case, to call attention to the fact that, say, a certain sensation is "generating" such a representational input is to step outside the purview of information processing and to begin what Haugeland calls "de-interpretation," "explanation by instantiation," and a kind of "nonintentional reduction." The sensation is an actual vehicle for the cognitive process. In this reductive step, (a component of) the cognitive process is "reduced" to a sensation. But this is irrelevant as far as the structure of the cognitive system is concerned and, thus, it is no concern of cognitivism. If this is true, then *both the intrinsic and the causal properties of sentient events are completely outside the purview of cognitivism.* Such events include, I maintain, not only pains and the sensory qualities, but also emotions, moods, essential components of aesthetic and even intellectual enjoyment and appreciation; in short, it is such events, states, and processes that comprise the most vital core of the "human spirit."

It seems to me, therefore, that Haugeland is wrong when he suggests that "there is no way to tell yet" whether cognitivism (the study of information processing systems) is ". . . at last the key to unlocking . . . [the mystery of the human spirit]." It is not, as he intimates, an "empirical" matter as to whether or not cognitivism can provide a comprehensive, exhaustive approach to all areas of psychology. Just as we knew (or should have known) all along that behaviorism could not provide such an approach (because we knew all along that we have rich "private, inner" experience – joys, sorrows, thoughts, pleasures, and pains), so we already know that a large, valuable portion of our experience is beyond the scope of cognitivism. Given the nature of our experience, this is not an empirical matter. It follows from the very limits that cognitivism sets for itself that much of the most vital purview of psychology is beyond these limits.

In spite of all of this, I still think it quite reasonable to believe that psychology (including the "human spirit" as part of its subject matter) can, in principle, be completely reduced to physics. It would be reduction of a somewhat special kind but still reduction in a clear, precise, all-out fashion. But perhaps I should warn that, in such a reduction, a portion of physics would, in a somewhat different though still quite similar sense, be reduced to psychology.

I take the subject matter of physics to be the manifold of physical events and the causal, spatiotemporal relations that hold among these events. To be a physical event is to occupy a position in the spatiotemporal, causal network. Brain events comprise a subclass of the class of physical events, and *to be a brain event is to occupy a position in an appropriate portion of the*

neurophysiological causal network. (I shall assume without argument that neurophysiology is, in principle, reducible to physics. Admittedly, this is not uncontroversial. It is questioned, I believe, by Fodor, 1978.) I now suggest the contingent hypothesis that mental events, in all of their mentalistic, qualitative, experiential richness, occupy positions in a certain appropriate portion of the neurophysiological causal network, that is, that every mental event is (also) a brain event. It would follow that mental events are (also) physical events and, therefore, part of the subject matter of physics.

It must be emphasized that I am not indulging in verbal trickery. It must be required for the reduction that mental events are discovered to be an integral part of the spatiotemporal, causal network that will be adopted by a "Utopian" or a "completed" physics. Whether or not this will be or would be the case is a contingent matter. (It is still another contingent matter, of course, as to whether or not we shall ever discover that such a reduction is, or is not, possible.)

Now, if the reduction is possible, all mental events are of course, physical events. But it also would follow, assuming that there are genuinely mental events, that some physical events are genuinely mental. This is what I meant when I said that a portion of physics would be "reduced" to psychology.

Is it absurd to hold that some genuinely physical events are also genuinely mental? Not, I believe, unless one has a hankering for old-fashioned, Hobbesian, everything-is-matter-in-motion materialism. Such a view is already rendered untenable by contemporary physics (to say nothing of introspection, even with "all of its ills," as Haugeland puts it).

But, I may still be asked, is it not absurd to try to "fit" mental events into the same spatiotemporal, causal network with forces, fields, electrons, electromagnetic quanta, and so forth? Surely, it may be objected, these are qualitatively in a different "category" from that one that contains mental phenomena. *This is precisely what I want to deny,* or, rather, I want to maintain that it is not necessarily the case. Physics, properly construed, has heretofore been concerned only with the spatiotemporal, causal structure of families of events. For example, a family of events with one kind of structure is called an "electron," one with another kind of structure a "train of electromagnetic radiation," and so forth. Properly construed, physics has been entirely noncommital concerning the qualitative or the instrinsic properties exemplified in individual events. This leaves open the possibility that some physical events (some brain events, we believe) *just are intrinsically* our pains, pleasures, thoughts, and so forth, as we live through them in all of their qualitative richness. (For a detailed explanation of this view, which was also advocated by Schlick, 1974, Russell, 1948, Feigl, 1961, and others, see my works, 1976 and 1978 forthcoming).

Let us consider a computer of whose inner composition we are ignorant. Suppose that after observing its inputs and outputs at some length we begin to theorize about its inner mechanisms. This could be done in several ways. Let us suppose that, in our case, we proceed by theorizing that the computer is composed of various units of "hardware." We then concern ourselves only with constructing theories about the causal interactions among these units. We are not concerned with what the individual units are; we remain noncommital about their chemical composition, their qualitative aspects, their intrinsic nature. There is a strong analogy between this kind of investigation and the kind of inquiry conducted by neurophysiologists and physicists in their study of the brain. Now suppose we expand the scope of our study of the computer. Either by further theorizing or by examining directly the interior mechanism, do we discover things about the "hardware." We find out what at least some of these internal units are; we learn something about their instrinsic nature. The corresponding analogy for the brain is simply that we develop and test neurophysiological and psychophysiological theories that provide information about the intrinsic nature of a select subset of brain events – information that these brain events are also mental events of certain specific kinds. Some of these will be sensations, "raw feels," and so forth, and, thus, will be at the noncognitive end of the spectrum. Some of these may even turn out to be "vehicles" for cognitive activity, as we saw above; and then there are also those mental events (that are also brain events) that are directly cognitive in nature. Insofar as we can put the cognitive activity of the brain in analogy with IPS computer activity, we must seriously countenance the possibility that some of the "hardware" involved is genuinely mental.

REFERENCES

Feigl, H. The Mental and the Physical. In: H. Feigl, M. Scriven, and G. Maxwell (eds.), *Minnesota Studies in the Philosophy of Science,* vol. 2, Minneapolis: University of Minnesota Press, 1961.

Fodor, J. A. Computation and Reduction. In: C. Wade Savage (ed.), *Minnesota Studies in the Philosophy of Science,* vol. 9, Minneapolis: University of Minnesota Press, 1978.

Maxwell, G. Scientific Results and the Mind-Brain Issue. In: G. Globus, G. Maxwell, and I. Savodnik (eds.), *Consciousness and the Brain.* New York: Plenum Press, 1976.

Rigid Designators and Mind-Brain Identity In: C. Wade Savage (ed.), *Minnesota Studies in the Philosophy of Science,* vol. 9, Minneapolis: University of Minnesota Press, 1978.

Russell, B. *Human Knowledge: Its Scope and Limits.* New York: Simon and Schuster, 1948.

Schlick, M. *General Theory of Knowledge.* Vienna and New York: Springer-Verlag, 1974.

by **John McCarthy**

Artificial Intelligence Laboratory, Stanford University, Stanford, Calif. 94305

Competence cognitivism vs. performance cognitivism. Haugeland's description of the attractions of cognitivism convinced me that I have long been a cognitivist, although I would not try to give cognitive explanations to all mental phenomena. His challenges to cognitivism are interesting, and here are some tentative ideas for meeting them.

We first distinguish *competence cognitivism* from *performance cognitivism.* Competence cognitivism tries to relate the output of a mental process to its inputs as the result of a deduction from its premises. Performance cognitivism regards the actual mechanism as a process of deduction. I would support performance cognitivism for some mental phenomena and competence cognitivism for many more. Arguments that the human brain obtains some of its results faster than can be accounted for by deductive processes are proper challenges to *performance cognitivism* but not to *competence cognitivism.* Thus we may cognitivistically relate the output of a skilled action to its input without postulating a deductive internal mechanism.

Performance cognitivism needs more precision, because it is not clear when a mechanism is to be considered cognitive, even in machines. In particular, pattern matching mechanisms are quite varied, and some of them are closer than others to deductive reasoning.

Here is a conjecture about "understanding." A person can deal with a symbolic input in either of two ways. First, he can manipulate it according to rules he has learned, and, second, he can translate it into his "internal language." Only in the latter case can he combine it freely with other information to draw conclusions. We say that someone has information but does not understand it when he can manipulate the symbols as objects but has not translated enough of it into his "internal language." The main symptom of lack of understanding is a failure to draw certain conclusions from symbolic information that seem "obvious" to those who do understand. This phenomenon offers no problems to cognitivism of either kind, provided we distinguish symbolic expressions from the information they express.

The most difficult of Haugeland's challenges to cognitivism is the problem of how moods affect reasoning. It is most acute if we imagine moods to be chemical; for example, a melancholy mood is just a high concentration of "melancholine" in the blood – together with its effects. How can this affect what long term goals seem worth pursuing, as moods often do? My proposals here are tentative and certainly require revision. The presence of melancholine might cause certain sources of sentences to come into consciousness with higher probability, or to present themselves with greater urgency, or to come with greater frequency. This assumes that the information retrieval system that determines what comes into consciousness is partly nonlogical. Likewise it inhibits other sources of sentences. The presence of melancholine might give rise to the sentence "I am melancholy." Perhaps this sentence has no special effect by itself, but can trigger learned effects. The sentence "I am melancholy" might be used logically by the information retrieval mechanism. Thus there might be rules saying something like "If I am melancholy, I should list things that might go wrong." This might work, but it seems subjectively implausible.

All these hypotheses require an explanation of a phenomenon that occurs in many other contexts. A certain set A of sentences can lead to a conclusion p. Another set B, consistent with A and maybe even containing A, can lead to the conclusion $\neg p$. McCarthy (1978) ascribes this to a method of conjectural reasoning we call circumscription induction. Briefly, applying circumscrip-

tion induction involves jumping to the conclusion that the objects whose existence follows from a set A of sentences are all the objects there are in a certain class. This conclusion may lead in turn to the conclusion p. Enlarging A to B may bring in new objects, and this may lead to $\neg p$. Without some such mechanism we would not be able to draw contradictory conclusions from consistent sets of information merely by neglecting part of it.

REFERENCE

McCarthy, J. *Circumscription Induction – A Way of Jumping to Conclusions*. Stanford Artificial Intelligence Laboratory, Stanford, University, Stanford, Calif., 1978.

by **Robert Monk**

Department of Philosophy, University of Illinois at Urbana-Champaign, Urbana, Ill. 61801

Cognitivism and cognitive psychology. Haugeland gives an abstract characterization of a position he calls "Cognitivism" and then gives reasons for doubting its eventual success. He does not clearly explain the relation between cognitivism and cognitive psychology. As I understand his view, cognitivism is a philosophical position underlying cognitive psychology, which is a branch of scientific research. He apparently holds that the philosophical position leads to a program of empirical and theoretical research committed to a certain sort of explanation of intelligent behavior. The research program, he thinks, will have great difficulty producing adequate explanations of the approved sort for certain phenomena that fall within its domain, notably moods, skills, and understanding.

Haugeland's paper is an excellent example of one kind of thing philosophers can usefully do for psychology: to explicate the philosophical underpinnings of important approaches in research. His accounts of explanation and reduction (instantiation) in cognitive psychology are masterful and convincing, and his careful analysis leading to definitions of "intentional black box" and "information processing system" is a model of good philosophical procedure. I find his defense of the two "cornerstones" of cognitive psychology persuasive, and also his defense of the view that cognitive psychology need not have the structure of physics in order to be a legitimate science.

I shall confine my critical remarks to the characterization of cognitivism and to the "hurdles." First, it is disappointing not to find a definition of cognitivism as clear and careful as those offered for "intentional black box" and "information processing system." To characterize it as the view that the mind is an IPS is suggestive but not really adequate. Although psychologists are beginning once again to use the term "mind" without embarrassment or contempt, it is not clear enough in meaning to figure in an analytical definition. It would also have been helpful to have been provided with a clear characterization of the relationship between cognitivism and cognitive psychology, a clear description of the domain of phenomena relevant to each, and an account of their explanatory obligations. Without these, the evaluation of the "hurdles" becomes more difficult.

Second, while it does seem to me that Haugeland's analysis captures the abstract form of an important class of theories in cognitive psychology, it also seems to me that cognitive psychology includes theories that are not cognitivistic in his sense. I have in mind image mediation theories, which are often advanced as explicit alternatives to theories using (quasi) linguistic coding. To give just one example, the work of Shepard and Metzler (1971, *op. cit.*) seems to support the view that subjects accomplish some mental tasks by performing mental operations with images, rather than by abstracting features and translating them into a code.

Third, not all cognitive theories that make use of quasi-linguistic coding look like models of a process of reasoning the problem through. Many simply exhibit a computer-like routine that will automatically produce the correct output for given inputs (Clark & Chase, 1972, present just one of many possible examples). Does running witlessly through such a routine constitute reasoning a problem through? In Haugeland's discussion of this aspect of an IPS I sense a confusion between generating the right outputs and generating a method for producing the right outputs. By his own definitions I should have thought that an IPS would simply produce outputs and not rationales for them. The rationale for the output would be built into the structure and operation of the IPS. We explain how an IPS arrives at its output for a given input by showing the intervening sequence of information-processing steps, and we explain the correctness or reasonableness of outputs by showing how a device constructed in that way must (barring malfunction, interference, etc.) produce correct or reasonable answers. The distinction between generating a method and generating outputs by that method is important, because it points to an important difference in research problems in cognitive psychology. There is the general problem of inventing models of routines by which human beings solve problems of specific types (such as three-term series problems), and there is the more fundamental problem of explaining how we come up with the routines we use, that is, how we put together the programs by means of which we solve these problems.

I find Haugeland's discussion of potentially serious hurdles for cognitivism the hardest part of his paper to grasp. I gather that the significant point about moods is that they can affect intelligent behavior and, therefore (by cognitivist lights), the outputs of IPSs, while not being components of or inputs into those IPSs. But there are many noncognitive states and processes that affect cognitive functioning: emotions, desires, motivation, deprivation, arousal, hormonal imbalances, fatigue, sleepiness, drug-induced states, tension, stress, shock, trauma, neurosis, psychosis, brain damage, birth defects, fever, nausea, disease, pain, and hiccups, to name a few. Why does Haugeland single out moods for special treatment? I am unable to follow the discussion of "segregation," but the principle Haugeland lays down (that segregated noncognitive states can be effective in determining intelligent behavior only insofar as they somehow generate quasi-linguistic representations that can be accepted as inputs by the cognitive IPS) seems arbitrary and much too severe. I do not see why a cognitivist could not maintain that noncognitive states and processes can simply disrupt the cognitive machinery in various ways, without there being inputs into the machinery concerning them. On the other hand, I agree with Haugeland (if this is one of his points) that noncognitive influences on intelligent behavior fall within the research domain of cognitive psychology, and that these are a potentially troublesome area.

The skills Haugeland mentions as he presents his prima facie reasons for doubting that the etiology of skillful behavior is cognitive are all conspicuously *motor* skills. While some cognitive processing is certainly involved in learning motor skills (the typist must read, the ping-pong player must follow the ball), it is indeed implausible (not to say absurd) to suppose that we learn them by exclusively cognitive means. But why this constitutes any sort of difficulty for cognitivism escapes me. Here the absence of an account of the explanatory obligations of cognitivism and its domain of relevant phenomena makes itself felt.

As Haugeland continues his discussion of skills, it begins to appear that what bothers him is that skills are often exercised without conscious thought. He implies that a theory of skills in terms of unconscious information processing could not be cognitivistic, because such processing would go much faster and be much more efficient than conscious thought. All of this comes as a great surprise. The stock in trade of cognitive psychology has been the explanation of skills and abilities that we exercise in a flash – recognition and identification of familiar patterns and objects, understanding speech, remembering familiar facts, solving trivial logic problems. These explanations have quite commonly been in the form of models that seem to fit Haugeland's requirements for being cognitivistic. There is nothing in his analysis to suggest that an IPS cannot work unconsciously, very fast, and very efficiently. So I simply cannot see the difficulty that skills are supposed to pose for cognitivism. It is also startling to find conscious thought presented as the archetype of cognitive processing. The cognitive movement in psychology has clearly taken as its principal archetype the computer routine, as Haugeland himself seems to acknowledge in his last section.

In his discussion of understanding, Haugeland is talking about very high level cognitive performance, which cognitive psychology is perhaps not ready to tackle. After all, we still do not understand how a person is able to recognize a dog or remember that all dogs are animals. It is difficult to judge whether "insight" constitutes a problem for cognitivism. A cognitive psychologist would scarcely be able to approach the topic without a clearer analysis of it and some experimentally demonstrated facts concerning it.

But whether or not there is an IPS explanation of insight, there may very well be some intelligent human performances that cognitive psychologists will be unable to explain by means of cognitivistic theory in Haugeland's sense. If so, what would be the consequences for cognitivism and cognitive psychology? Haugeland suggests (but never says) that the two stand or fall together. This is surely not so. Cognitive psychologists are free to invent

modes of theorizing and conceptions of human information processing that do not conform to the cognitivist paradigm. (Image processing is apparently one such mode; holographic processing may be another.) If some intelligent behavior requires such innovative explanations, then cognitivism will have to be abandoned as the exclusive account of cognitive processes. That would not necessarily mean that all cognitivistic explanations must be scrapped, for they work where they work. And it would not mean the death of cognitive psychology, but only of one perhaps rather narrow and confining conception of what, in the abstract, cognitive psychology is supposed to do.

REFERENCE

Clark, H. H., and Chase, W. G. On the process of comparing sentences against pictures. *Cognitive Psychology*. 3:472–517, 1972.

by **Thomas Natsoulas**

Department of Psychology, University of California, Davis, Calif. 95616

Haugeland's first hurdle. Moods may indeed pose a problem for cognitivist explanation. Moods are general states of consciousness (certain ways of being conscious, certain chronic modifications of one's consciousness) during which, typically, a great variety of conscious mental episodes (perceptions, thoughts, etc.) occur. [At work here are two senses of consciousness: (1) A conscious mental episode is one that is *conscious*$_4$ (see the numbered entries under "consciousness" in the *Oxford English Dictionary*) It is, or can readily become, an object of direct awareness such that the person whose mental episode it is becomes, noninferentially, aware of its occurrence. (2) The other uses of "conscious" and "consciousness," in the preparenthetic sentence pertain to *consciousness*$_6$. The Dictionary defines consciousness$_6$ as "the state of being conscious, regarded as the normal condition of healthy waking life." To be conscious$_6$ is "to have one's mental faculties actually in an active and waking state." While we are conscious$_6$, we pass through moods, which are certain ways of being conscious$_6$. They are modifications of our consciousness$_6$, though they are not departures from it.] Haugeland suggests that the problem about moods lies in the evidently widespread influence a mood can have on which mental episodes occur while it holds sway; how these effects come about does not seem to him to be open to intentional interpretation.

Moods do not have their pervasive effects as a consequence of our being conscious of their occurrence (although, of course, we are often aware of being cheerful, melancholy, etc.). No view is proffered here, however, on how moods function to produce their far-reaching effects. Instead, it is questioned whether cognitivism requires that all noncognitive causes of intelligent behavior be mediated by representations of those causes, for this requirement is supposed to be a source of the difficulty about moods.

The aim of cognitivist explanation is to account for the intelligent behavior of objects (or organizations of objects) construed as information-processing systems. These systems consist of concrete intentional black boxes as functional components that interact to produce the outputs of the system. The inputs and outputs of the system and of the component black boxes are said by Haugeland to possess *intrinsic properties* (a) that are *not* included in the intentional interpretation of the input-output relations, and (b) whose existence removes any mystery about how intentionally interpreted states might be in the world. As objects in nature, information-processing systems are potentially subject to myriad noncognitive influences; some of which surely affect intelligent behavior. Think of the effects drugs called "consciousness-expanding" can have on cognitive processes.

Cognitivist explanation of such influences on intelligent behavior is not ruled out by their noncognitive character. Presumably, they have their effects on intelligent behavior by modifying the function and interaction of the intentional black boxes that comprise the particular information-processing system. Thus, the new input-output patterns (when under the influence) would be explained systematically and without a "change of dimension." Cognitivist theories, *which presuppose certain regularities,* would be formulated with regard to alternative input-output patterns of the various functional components and alternative interactions among them.

We know that some information-processing systems have different possible settings; they have the ability to function in different modes. In other words, the pattern of their input-output relationships is subject to systematic variation through determinate and reversible changes in how their components function and interact. Perhaps a change in mood brings about a change in mode in this sense.

The regularities presupposed by cognitivist explanation of this kind are not all interpretable intentionally; certain noncognitive influences are not mediated by representations of those influences, or by other (literal) inputs. These regularities must be treated as unexplained, as long as one remains within the framework of cognitivist science. They pertain to certain variable noncognitive conditions of cognitive functioning in the world.

How Haugeland expresses his view of moods as a serious hurdle for cognitivism suggests that he may have in mind another characteristic of moods (in addition to their pervasive influence). He writes of them as affecting *and permeating* all kinds of cognitive states and processes. Perhaps one cannot treat them merely as causes because they "infect" the conscious mental episodes they affect and "transform their meanings." When one is melancholy, everything seems greyer, duller, and less livable; one might have said that moods "color" our perceptions, thoughts, memories, expectations, and so forth. Haugeland seems on the verge of proposing that a mood provides conscious mental episodes, which occur while the mood holds sway, with a certain qualitative character, and that this qualitative character is ignored by cognitivist explanation, which interprets all the component inputs and outputs of any information-processing system as quasi-linguistic representations.

In a less obscure context, Haugeland does mention this sort of objection to cognitivism, but only to dismiss the objection. Others suppose it is a problem that certain conscious mental episodes (e.g., visual awarenesses) have a qualitative character that plays no role in cognitivist explanation. But Haugeland assumes all such mental episodes are noncognitive causes of inputs that can be interpreted as quasi-linguistic representations (e.g., a certain noncognitive sense-impression produces an episode of seeing what color the grass is).

Once we recognize that particular conscious mental episodes are often *both* cognitive and qualitative, the problem for cognitivism reasserts itself. For example, seeing that the grass over there is green involves (a) the occurrence of a quasi-linguistic representation (greenness predicated of the grass) and also (b) the qualitative presence of the grass. In other words, the grass is not just known or thought to be green; there is also its concrete perceptual (i.e., visual) presence, its specific qualitative presence in the visual modality. The mental episode would not be an episode of seeing without its qualitative character.

It needs to be shown that a component input's (or output's) being qualitative makes (or does not make) a difference to the explanation of intelligent behavior. Qualitative character may be a matter of intrinsic properties independent of intentional interpretation. Or qualitative character may be implicit in the meanings assigned to certain inputs and outputs by cognitivist explanation. That is, an interpretation of input along the lines of "this grass is green" may suffice as part of an explanation for certain intelligent behavior (e.g., snapping a picture of grass) *only because* it is taken to mean that the grass was qualitatively present in the visual way.

Why might reference to qualitative character be necessary for the explanation of intelligent behavior? The reason can only be sketchily indicated here. A degree of perceptual determinateness is often required by intelligent behavior, a degree that cannot be provided in most cases by the conceptual aspect of a perception. In such cases, it is by virtue of the qualitative presence of the object or scene (to be behaved towards) that the respective mental episode's reference is fixed. Perceptions, as quasi-linguistic representations, would often be assigned a meaning in terms of demonstrative constructions. As in the parallel linguistic case, in which full interpretation requires an extralinguistic act, the interpretation of perceptual inputs must include the extraquasi-linguistic dimension of qualitative presence in order to tie the perceptual reference down to a specific environmental cause of the input.

by **Herbert R. Otto**

Department of Philosophy, Plymouth State College, University System of New Hampshire, Plymouth N.H. 03264

A program is not an explanation. Systematic explanation differs from derivational-nomological in that the former, but not the latter, necessarily appeals to "specific structures" and the "abilities or dispositions" possessed by these structures. On the other hand, it differs from morphological explana-

tion by the reference it must make to a "complexly organized pattern of interdependent interactions" (section 2). Granting such distinctions, one can argue, as Haugeland does, that the problem of "significance" or "meaningfulness," that is, the problem of "connecting 'meanings' with the physical order of cause and effect," can be circumvented (section 3). This is done by locating, in the subject matter, interpreted functional units (IBBs or "intentional black boxes") and then, without "de-interpreting," providing a systematic explanation of the workings of such units. Where this can be done, the set of such units is called an information processing system (IPS). The great advantage of Cognitivism then, is that because of this unique methodology, we are able validly to regard the mind *as an IPS*. Hence, "intelligent behavior is to be explained by appeal to internal 'cognitive processes'. . . interpretable as working out a rationale" (section 4).

Now, what objection do I have to all this? Certainly it is not to the general thrust, for I am impressed with Cognitivism and its potential. Rather, I am uneasy about certain premises that seem to me not only to render the argument inadequate to its task, but that could be seriously misleading as well. First, I am not yet convinced that there are, in fact, distinct "styles" of scientific explanation. From Haugeland's examples, I can conclude no more than that there are differences of *complexity* among various explanations. By this I simply mean that when properly formulated an *explanans* may involve a chain of more than one deductive argument, and when it does, the "laws" utilized as premises in the component arguments need not all be drawn from the same scientific discipline. Thus, unified science is not ruled out merely on logical grounds, nor are we led to make metaphysical presuppositions about subject matters. The paradigm shift is not one of explanatory style, but rather one of change of view regarding the "rational agent." Cognitivism firmly rejects the old view of the mind as passive (Locke's *tabula rasa*) in favor of a dynamic model. And, thus, although Cognitivism's underlying pattern of explanation need not differ from that of the other sciences, it is required that new symbolic means with which to express this dynamic aspect be devised. As I see it, this is exactly what the program flow chart, with its logic of command, feedback, and conditional branching, provides. It is a new species of mathematics. The computer then sets the empirical constraints (see Pylyshyn et al, BBS 1:1,1978) and provides for testing consistency and completeness. But the program is not itself an explanation; rather, it is a concrete application of one. To illustrate: a chess program is to an explanation of chess-playing behavior as, say, a prototype of an airfoil is to an explanation of aerodynamic lift.

by **Steven Pinker**

Department of Psychology and Social Relations, Harvard University, Cambridge, Mass. 02138

Mind and brain revisited: forestalling the doom of cognitivism. Haugeland argues that if cognitivism fails as a suitable metatheory for psychology, the problem of moods might be its Achilles' heel. I would like to examine a class of problems that may offer similar headaches for cognitive theorists, and suggest two possibilities that imply the existence of reasonable solutions within a cognitivist framework and a third that seems far less tractable.

Why should moods seem so problematic to Haugeland? Surely it is not because a cognitivist account of their effects on thought processes is impossible. After all, the cognitivist's bag of explanatory tricks contains the full power of a Turing Machine, and it would be a routine (if tedious) matter to add boxes and arrows to information-flow diagrams that would suitably readjust the estimates of subjective probabilities, assignments of salience to stimuli, affective evaluations of other people, and so forth. What must concern Haugeland, and legitimately so, is the possibility that the effects of moods cannot be incorporated into a cognitivist theory of thought in a principled or elegant way. Similar doubts can be raised (although Haugeland does not invoke them) concerning the effects of development, senility, drug states, schizophrenia, and brain damage. What these cases share is that thinking seems to be affected in a drastic, pervasive, and qualitative manner, physiological factors are or may be implicated, and elegant cognitive accounts seem elusive. I hope to show that the latter two conditions may be intimately related.

As Haugeland notes, most investigators assume that a bridge between physiology and cognition exists at the level of the lowest or smallest units of cognitive explanation. This is an obvious parallel to the way that computer programs can be written in high-level languages, themselves programmed in lower-level languages, and so on down to the level of the machine language that is physically instantiated in the circuitry of the computer. For example, Simon and Newell (1964) speak of a small set of "elementary information processes" as the atoms of psychological theories and the fundamental mechanisms that physiologists must explain; Scott (1977) outlines seven levels of explanation for brain science, levels five, six, and seven of which are (respectively) the elements of thought from neural network dynamics, psychology from the elements of thought, and human personality and culture from psychology; and Dennett (unpublished) describes mental organization as a hierarchical committee of homunculi, the stupidest of which can be "replaced by machines." Part of the appeal of this scheme is that it allows cognitive psychology to proceed with a policy of benign neglect toward physiology, since the "units of thought" have a particular function in higher-level theories regardless of the exact nature of their physical instantiation, or, as Haugeland notes, whether they have any physical instantiation at all.

Given that this scheme is correct, there are two ways of accounting for extreme qualitative variations in thought such as moods, insanity, development, and so forth. One would be to assume that physiology plays no direct role at all: I have already suggested what such an explanation for moods might look like. In the case of development, we could imagine that the human infant is "bootstrapped" with an extremely powerful and intricate program that interacts with its environment and seems to change in complex ways that are similar in principle to the chess-playing programs that learn from their mistakes, or to a theorem-proving program that uses its solutions to one problem as a component of its solution to another.

A possible objection to this move is that the minds of a six-month-old and a twenty-year-old, or of a normal person and a schizophrenic seem so radically *different* that it strains credulity to say that a single master program might be behind them all. If one insisted on describing such a program in detail, it might look too much like a Rube Goldberg contraption to be plausible; but if one described the program in a general way, it might be inadequate to account for the phenomena. The best example of the latter horn of the dilemma is Piaget's claim (see *oper. cit.* in Brainerd et al., this issue) that the child's internal representational system increases in logical power as a result of certain interactions with the environment that are mediated by that representational system. As Fodor (1975) and Macnamara (1976) argue, Piaget cannot have it both ways: either the full logical power is in the representational system to begin with, or the change is precipitated by factors external to the representational system itself.

This leads to a second possiblity, that qualitative cognitive effects, such as the elementary information processes or the buffers and memories they use, can be brought about by quantitative physiological causes via the bridge between physiology and cognition. Thus, moods might come on because of a reallocation of processing capacity to different components, schizophrenics may have deficient attentional filters, and children may develop because their short-term memory increases, allowing more plans to be executed at once (Pascual-Leone, 1970). From here, the baton is passed to the physiologists, who must explain the substrata of the attentional deficit or memory growth. This scheme has some plausibility; automata theory shows that one can delineate classes of algebras, logics, or languages simply by constraining the size or accessibility of the working memory of the machines on which the structures can be instantiated. Whether this will work in the desired way for more complex systems is less certain. Perhaps small changes at the bridging level can percolate upward through the different levels of organization until they give the behavior of the system new and emergent properties; on the other hand, in many instances the system might simply "crash." In any case, we have a second way that cognitivism might handle Haugeland's type of objection.

However, if both of these routes go nowhere, a third and more ominous possibility exists. It may be that the physiological properties of the nervous system are tied to cognitive processes at many levels, not just the physical instantiation of the atoms of information processing. For example, one could install in a computer hard-wire microprocessors or analog subsystems (say, a differential analyzer) that could perform extremely complex computations without being decomposable into machine language instructions or any other interlingua. An observer unaware of such a modification might be hard pressed to explain the changes in the computer's behavior in terms of a quirk in its program or a simple alteration of its buffers or registers. In the case of human beings, physiological maturation, a bad gene, or the sudden release

of a neurotransmitter could impinge on hundreds of thousands of individual neurons in highly specific ways and dramatically affect thought and behavior. Again, attempts to explain the changes in terms of accommodations of cognitive schemas or small adjustments of memory or attentional mechanisms would be fruitless. If this is true of moods, development, and so forth, then, indeed, Haugeland is correct in warning that an elegant and purely cognitive explanation may not exist. This would not necessarily spell the doom of cognitivism, however, just of cognitive isolationism – cognitive psychology and physiology would not be islands linked by a two-lane bridge, but countries sharing a long border. Cognitive theorists could ignore physiology only at their peril in explaining deviations from normal, cool-headed adult cogitation. However, since this complicates the task of theory construction tremendously (there are many more degrees of freedom), it seems a reasonable strategy for researchers to pursue the first two options, pretending that the third does not exist, for as long as possible.

ACKNOWLEDGMENT

I am grateful to Stephen Kosslyn for his helpful suggestions.

REFERENCES

Fodor, J. *The Language of Thought,* New York: Thomas Crowell, 1975.

Macnamara, J. Stomachs assimilate and accommodate, don't they? *Canadian Psychological Review.* 17:167–73. 1976.

Pascual-Leone, J. A mathematical model for the transition rule in Piaget's developmental states. *Acta Psychologica.* 32:301–45. 1970.

Scott, A. Neurodynamics: a critical survey. *Journal of Mathematical Psychology.* 15:1–45. 1977.

Simon, H., and Newell, A. Information processing in computer and man. *American Scientist.* 52:281–300. 1964.

by **Karl H. Pribram**

Department of Psychology, Stanford University, Stanford, California 94305

Image, information, and fast Fourier transforms. What a refreshingly clear view of cognitivism Haugeland presents in this paper. He does not dismiss the cognitive approach as trivial (though it might turn out that way) nor as a panacea for all of psychology's problems. There are, therefore, only a few comments that I venture.

Haugeland makes a good case for separating holographic image-processing from information processing – a distinction that I had not made so clearly for myself. Miller, Galanter, and I have discussed (1960) Image and Plan as mutually implying each other. In Haugeland's terms, image-processing and information-processing are different dimensions, and cognitive scientists need to formulate the rules that allow translation from one to the other and back again. I believe this is the contribution Yevick (1975 *op. cit.*) is attempting to make, and it is thus more important than just a demonstration that a holographically based "logic" can be constructed.

Some of the character of translation rules necessary to Haugeland's "intentional interpretation" of the relationship between image- and information-processing can be gauged from a model such as that proposed by Luce and Green (1972). Their physical instantiation is with regard to patterns of neural impulses: the departure of interresponse histograms from a Poisson distribution. But in order to perform the necessary manipulations of data, the authors resort to computer-processing and specifically to using fast Fourier transforms (FFT) for convenience. But holography consists of encoding in the frequency domain, and a physical instantiation of this part of the Luce-Green model might suggest that the FFT is performed at synaptic junctions. This is a suggestion that can be supported by a considerable body of data (Pribram, 1971; Pribram, Nuwer, & Baron, 1974 *op. cit.*). The translation rule from interresponse time to the frequency domain is thus straightforward. The inverse Fourier transform provides the computer readout in digital terms, but in this case the translation rule is not so easily instantiated in the physical nervous system. Furthermore, other parts of the Luce-Green model refer to other processes such as "decisions" that may or may not have anything to do with holography. The character of any translation of the information-processing aspect of the model and the holographic (FFT aspects) is therefore not just the invertibility of the transform (though this is an important component), but the complex of steps in which the transformation is (or is not) resorted to. As Haugeland notes: "Who knows what could happen among holograms in intersecting planes or . . . scattered about in volumes – especially if they dynamically modified one another in non-chaotic ways? Perhaps we will find there the translation rules that will bring image processing and information processing together into a more complete cognitive perspective." Haugeland's analysis should help us considerably to know whether attempts at such a synthesis are spurious or sound.

REFERENCES

Luce, R. D., and Green, D. M. A neural timing theory for response times and psychophysics of intensity. *Psychological Review.* 79 (1):14–57. 1972.

Miller, G. A., Galanter, E., & Pribram, K. H. *Plans and the Structure of Behavior.* New York: Holt, Rinehart & Winston, 1960.

Pribram, K. H. *Languages of the Brain: Experimental Paradoxes and Principles in Neuropsychology.* Englewood Cliffs, N.J., Prentice-Hall, Inc. 1971.

by **Roland Puccetti**

Department of Philosophy, Dalhousie University, Halifax, N.S., Canada B3H 3JS

Are right hemisphere activities cognitivistic? Near the end of his paper Haugeland considers the problem posed for cognitivism by *skills.* The greater the degree to which one has mastered a skill, the less its exercise seems to depend on internal cognitive processes. As he says, this is evidenced by our inability to articulate how we do it, by the fact that thought and deliberation cease at just about the time we come to do it well, and by the fact that thereafter when we think about what we are doing performance declines. Haugeland says a cognitivist might try to "explain away" these phenomena by postulating an unconscious information processor at work in executing skills, but suggests this would be very ad hoc.

However, the examples he gives up to this point – playing ping-pong, typing, and piano-playing – all involve motor skills, and probably the storing and running off of such learned movements does depend on unconscious encephalic mechanisms, in either the cerebellum or the midbrain.

Quite another matter is the question of chess skills, which Haugeland considers next, and where also he thinks cognitivism might resort to bringing in the notion of an unconscious information processor in order to save appearances. But then, Haugeland says, cognitivists would have to face Dreyfus' (1975 *op. cit.*) interesting pointed question, which he recasts as follows:

"It is known that intermediate, advanced, and great chess players are alike in consciously considering on the order of a hundred plays in thinking out a move; they differ in their 'skill in problem conception' (de Groot, 1965) – i.e. in preselecting which moves to think about. Now the rationale for these good preselections would be enormously long if they were spelled out (many thousands of plays). It's possible that players have some marvelously efficient unconscious information processor which works through these rationales; but if so, then why would anyone with such a splendid unconscious ever bother to deliberate consciously and tediously over a hundred plays? The implication is that the skillful preselection and the tedious cogitation differ not just in efficiency and consciousness, but in kind, and that neither could adequately substitute for the other. [section seven]."

But it does not follow from the chess player's being unconscious of how the moves he studies in depth were preselected that this preselection was done unconsciously. For as I have argued (1974), it may just be that despite our intuitions, not only the preselection but the insights necessary for successful cogitation over the moves one studies in depth are supplied across the forebrain commissures by the mute (usually the right) cerebral hemisphere. Nor is it necessary to suppose that the preselection involves running through thousands of possible continuations to select the promising ones, as a digital computer would have to do; the so-called minor hemisphere may simply see what are attractive lines of development, the way a professional painter or sculptor can quickly pick out in a huge amateur art exhibit the few pieces showing talent. It is able to do this, I suggest, because the characteristic cognitive mode of the nondominant (for speech and handedness) hemisphere is not analytic and sequential but synthetic, holistic or Gestalt, precisely what one would expect if it is specialized for visuospatial perception.

Underlying the above argument, of course, is the equally radical assumption that chess-playing is not a matter of deductive inference but of perceiving dynamic interrelationships on the board, hence requiring not logic so much as visual imagery. In this connection I refer to another study, by de Groot (1966), where he found that a five-second exposure to a complicated

midgame situation was sufficient for a master, though not for a player of expert rank, to reproduce the situation on another board with few or no errors; yet when the pieces were placed at random, so they no longer formed a meaningful *chess pattern*, masters did no better than beginners at reproducing the position. Thus the model for chess appears to be not mathematics in general but geometry.

Evidence that it is indeed the nonspeaking hemisphere that is dominant for processing geometrical-type cognitive tasks now seems unassailable. Franco and Sperry (1977) have recently reported a study in which commissurotomy subjects were asked to match objects that they could touch but not see with geometrical shapes they could see. The superiority of the left hand (controlled by the right hemisphere) over the right hand (under control of the left hemisphere) in these right-handed people was consistent throughout, and increased dramatically as the type of geometrical task progressed from more to less structural constraints involving Euclidean, affine, projective, and topological space. Indeed, the disconnected left hemisphere, using the right hand, performs barely above chance level on topological sets. Normal controls show no significant differences in hand performance on the same tasks, indicating that the speech hemisphere is utilizing spatial perception provided by its mute companion through the intact commissural connections. The authors comment (p. 112):

"The present evidence strongly suggests that . . . preverbal apprehension of geometrical relations is mainly a right hemisphere function. It seems therefore likely that right hemisphere operations are primary in the apprehension of geometrical properties of space and that these only subsequently become susceptible to verbalization. An active interaction of the two hemispheres during processing of geometrical problems seems implied and a similar interhemispheric integration would seem reasonable also for other cerebral activities involving spatial intuitions and their linguistic expression."

Applying this now to chess, I am suggesting that the right cerebral hemisphere processes board situations consciously, but that we who talk and write and read and calculate are not directly aware of this processing, let alone how exactly it is done. Thus, if this is correct, cognitivism need not introduce the notion of an *un*conscious information processor at work on a task that is, after all, an example *par excellence* of intelligent behavior.

But that does not mean cognitivism is off the hook where chess skills are concerned – far from it. For, as I have said before (1975), those who take the computer as a model of cognitive information processing will be hard put to assimilate right hemisphere functions to that model. (How will we ever, for example, devise right hemisphere-type algorithms for a computer?) Thus Haugeland and Dreyfus may be on the right track here, even if they have got the mechanism wrong. Cognitivists have a lot left to explain.

REFERENCES

de Groot, A. D. Perception and Memory Versus Thought: Some Old Ideas and Recent Findings. In: B. Kleinmuntz (ed.), *Problem Solving.* New York: Kreiger, 1966.

Franco, L., and Sperry, R. W. Hemisphere Lateralization for Cognitive Processing of Geometry. *Neuropsychologia.* 15:107–13. 1977.

Puccetti, R. Pattern Recognition in Computers and the Human Brain: With Special Application to Chess Playing Machines. *British Journal for the Philosophy of Science.* 25:137–54. 1974.

Discussion. In: S. R. Harnad, H. D. Steklis and J. B. Lancaster (eds.), *Origins and Evolution of Language and Speech. Annals of the New York Academy of Sciences,* 280:262–64. 1976.

by **Georges Rey**

Department of Philosophy, State University of New York, Purchase, N.Y. 10577

Worries about Haugeland's worries. I am sympathetic to Haugeland's sympathies with cognitivism in psychology. Cognitivism does seem to promise an account of many human and animal states and behavior, one that is "rigorous and empirical," while, unlike behaviorism, its predecessor in this line, at the same time preserving many of our antecedent mentalistic intuitions. It seems to offer what Haugeland has aptly described as a "systematic," and, I would add, naturalistic framework within which to accommodate many of our beliefs about, for example, our own beliefs and desires and their causes and effects in the world around us.

But I am worried about Haugeland's worries. He sees cognitivism as having to confront three "serious hurdles": "moods, skills and understanding." Comparing his descriptions of these hurdles with his characterization of cognitivism, however, it is very difficult to see just what is so serious about them. They seem to have little or nothing to do with the cognitivism Haugeland describes. It is worth making the comparison in considerable detail, thereby we hope, assuaging Haugeland's, and perhaps many others', fears.

Consider skills. Haugeland claims (1) that they are independent of "articulateness" about them; (2) that "thought and deliberation" are in fact often excluded by their exercise; and (3) that their exercise is often "faster than thought." (By "thought" he appears to mean – as, on pain of begging the question, he must mean – *conscious* thought.) These claims all seem plausible enough. Haugeland asserts that they constitute "three *prima facie* . . . reasons for doubting that the etiology of skillful behavior is cognitive." But, prima facie as they may be, they seem to have nothing whatever to do with cognitivism as Haugeland has defined it. "The fundamental idea of Cognitive Psychology," Haugeland writes, is that "intelligent behavior is to be explained by appeal to internal 'cognitive processes' " – meaning, essentially, processes interpretable as working out a rationale. Cognitivism, then, can be summed up in a slogan: "the mind is to be understood as an IPS." Understanding, Haugeland's chess-playing Black Box as an IPS; for example, "means there is a systematic explanation of how it manages to come up with legal and plausible moves as such, regardless of how it manages to . . . do whatever it does that gets interpreted as those moves." Surely he might have added: "and regardless of whether it deliberates, has conscious thoughts, or is capable of providing, as further output, representations of how it manages to come up with those moves." At any rate, no such conditions of "deliberation," "conscious thought," or "articulateness" are included in his definitions either of cognitivism, of an IPS, or of an IBB. And would it not be odd to add them? In the first place, his very paradigm of an IPS, that very Black Box, would directly fail to satisfy them (or is it, and Minsky's machines at M.I.T., to be regarded as somehow conscious, and all that goes with it? If they are, then so might we be, even when we would least suspect it.) In the second place, such conditions would seem to have precious little to do with the regularities cognitivism is attempting "systematically" to explain. Inputs and outputs can presumably "make reasonable sense" whether or not the system also engages in self-monitoring. Indeed, I would have thought that part of the empirical respectability of cognitivism derived from its focus upon these sense-making regularities, instead of depending upon "the supposedly disreputable method of introspection."

In the light of all this, it is thus especially disconcerting to notice Haugeland's scorn for the "unconscious." He does argue, quite plausibly, that "skillful [unconscious] pre-selection and . . . tedious [conscious] cogitation differ not just in efficiency and consciousness, but in kind": there certainly do seem to be all sorts of dramatic differences between conscious and nonconscious processes. But he has provided no reason whatsoever for thinking this difference in kind coincides with the difference between the cognitive and the noncognitive. On the contrary, the examples of his Black Box, modern computers, many animal behaviors, and most human ones (e.g., understanding natural languages, working out unconscious motives) all suggest that the difference is one that ought to be drawn *within* the realm of the cognitive. (And, indeed, one might rightly demand of a cognitive theory that it explain just what it is about self-monitoring and conscious cogitation that makes them, while inefficient, so useful. I see no reason why such a theory could not meet that demand.)

Put another way: it is not that cognitive processes are, as they have been traditionally assumed to be, prima facie conscious, with their occasional unconsciousness calling for the explanation. Rather, as Haugeland's and others' accounts of cognitivism suggest, cognitive processes per se need not involve any consciousness at all: "consciousness," if it is a trait of anything, is a trait of only *some* states of only *some*, very special cognoscenti, some particularly complex IPSes, namely, human beings. At any rate, it seems to be only human beings that act with "thought" and "deliberation" and are able to report on *some* – although, as we know just from Freud, by no means *all* – of their cognitive states. What needs explaining is not our *unconsciousness,* but rather, how in the world we manage to do such remarkable things as "deliberate," "consciously cogitate," and be so "articulate" about ourselves, and how those things manage, sometimes, to affect the rest of our states and behaviors.

A similar discrepancy between Haugeland's "hurdles" and his account of cognitivism occurs when he worries about moods. He makes quite a number

of claims about moods. Some of them are far from obvious: I do not see that we are yet in a position to know whether, or to what extent, moods are "inferential," "quasi-linguistic," or "rational." Some of them are simply false: we constantly justify our melancholy by citing our misfortunes, or our joy by citing our good luck. And recent work on emotions suggest that at least they are quite highly cognitive. (See Schachter & Singer, 1962, pp. 380, 398, who present surprising evidence for their claim that "an emotional state may be considered a function of a state of physiological arousal and of a cognition appropriate to this state of arousal.") At any rate, Haugeland's claim that moods "don't seem at all cognitive" seems just groundless and extravagant.

But even supposing that, oddly enough, moods turn out to be not at all cognitive, still it is difficult to see how they would therefore constitute a hurdle to Haugeland's cognitivism. If moods are *not* part of our "intelligent" behavior, as Haugeland seems to be claiming, then, again by his own definition, cognitivism does not even *aspire* to explain them. Digestion, too, undoubtedly affects and is affected by cognitive states, while not being itself cognitive. But surely that is hardly a hurdle for cognitivism!

Moods quite probably lie in an area of psychology that straddles both cognitive theory and physiology. This is an area that, rather than presenting a hurdle for cognitivism, is, on reasonable assumptions, demanded by it. Any purveyor of an IPS explanation – whether cognitive psychologist or computer scientist – expects his IPS program to be realized in some mechanical medium or other (indeed, it is the fact that it is relatively clear just how an IPS program can be mechanically realized (cf. Section five) that *pace* Section six) lends cognitivism its naturalistic plausibility). In being so realized, whether in transistors or in quivering protoplasm, there will inevitably be properties of the realizing material that, while being themselves inessential, will interact with properties that are essential to the IPS program (consider, for example, the importance of the time neural transmissions take). These further properties will therefore be necessary to a full account of the psychology of the actual, spatiotemporal cognoscentum in question. It follows that the cognitive theory alone would not be sufficient. People are embodied minds; moods seem to involve properties of both that mind and that embodiment, and an account of them involves, therefore, both cognitive theory and physiology. But that a phenomenon involves both cognitive theory and physiology is an argument for, not against, cognitivism. From the fact that cognitive theory is not sufficient as psychology, it does not follow that it is not necessary.

Lastly, Haugeland is worried about "insight" and "understanding." He fears that a successful cognitivist theory would "preclude any radically new ways of understanding things." Now, it is certainly true that a successful cognitive theory would tell us a great deal more than we presently know about the constraints we impose upon acceptable, "intelligible" theories: on what "makes sense." Perhaps, as Haugeland seems to suggest, we not only have constraints upon theories, but constraints upon constraints upon theories: "general cogency conditions." It is not easy to focus on precisely what suggestion Haugeland does have in mind in his discussion of "insight." The word is introduced to distinguish rote from more conceptual understanding of "what's going on;" but, a few sentences later, it is identified with a more *meta*-theoretical "ability to tell when a whole account . . . makes sense"; and, still later, it seems to involve a *meta-meta*-theoretical recognition that "certain new constraints [upon acceptable theories?] constitute a kind of cogency." Only if all these different levels are run together is it plausible that "there could be a rationale [at the highest level] only if the new conditions were equivalent to, or a special case of the established ones." I see every reason for keeping the different levels distinct. However, no matter how abstract and meta-theoretical the ultimate cogency conditions may turn out to be, there is still all the difference in the world between knowing those ultimate conditions and knowing the results of applying those conditions to their respective domains. It is, after all, one thing to understand, or even construct, a computer program to solve some particular problem, and quite another to solve the problem oneself. Consider again Haugeland's chess-playing Black Box: we might well come to understand how the box sorts and weighs different chess positions, how it constrains its moves, without our being able to employ effectively these sortings, weightings, or constraints ourselves (it might well be that our applying the Box's procedures through the medium of our language and "conscious thought" would be inordinately cumbersome and inefficient). Thus, for all our "insight" into the Box, it might well provide us ever newer "insights" into chess. Similarly, if we ever did succeed someday in understanding the principles Richard Feynman's mind employs in selecting plausible physical theories – or in selecting plausible principles for selecting plausible theories; or, for some n, in selecting plausible n-order principles for selecting plausible (n – 1)-order principles – that would be a far cry from ourselves thinking of those theories, or of those (n – 1)-order principles We just might not be as "smart" as he is: we may not employ the principles as well or we may not generate as imaginatively the elements of the domain to which the principles apply; probably both are true. And so Feynman might well stun us, as did Galileo, Kepler, and Newton before him, with "a totally new way of talking about what happens," or even (depending upon how abstract the operant domain is) "a new way of making it intelligible." Understanding merely understanding may leave a great deal still to be understood.

Haugeland is skeptical about isolating the "ultimate general cogency conditions." And perhaps one should not quarrel with a healthy skepticism. But one should quarrel with weak reasons that nourish it: the fact that "people who regularly make convergent decisions about the reasonableness of theories and interpretations don't explicitly work through rationales . . ." we have already seen to be quite irrelevant to the project of supplying a cognitive account of those decisions. And the fact that "barrels of philosophical ink have been spilt," attempting to make sense of "making sense," "so far without success" is just false: try the astonishing advances in logic and the theory of computation that we have witnessed in the last hundred years! It seems to me that a good deal more success would be realized in this area were we not to regard the project so monolithically. The ordinary notion of "making sense," to none other than which Haugeland must be appealing, involves a variety of very different relations, each of which quite probably requires separate treatment: there are not only the deductive relations explored in logic, but also what might be called the "practical" relations (among beliefs, utilities, and actions) explored in decision theory, and the theory of action; as well as the "inductive relations" (between evidence and "the best explanation" of it) explored in inductive logic and the philosophy of science. And still more success might be possible were syntactic sense-making relations (relations among the representations themselves) treated separately from semantic ones (relations between representations and what they represent); and were it remembered that it is not only the relations between input and output that are important to the cognitivist, but also the relations between (and the quasi-linguistic character of) the intervening states, which break the syntactic relations down. Perhaps some of the sources of Haugeland's despair about success in this area can be traced to his failure to consider any of these distinctions. That the work in this area is far from finished is due in part to the fact that a plausible way of going about it was not really available until these recent advances were realized; indeed, the very idea of a cognitive psychology was a result of them.

I suppose the reason for postulating some not-all-that "mysterious and magnificent," but, to be sure, largely "unconscious," IPS is not merely that it would "facilitate a tidy account of intelligence," but that it offers the best promise so far: we can imagine, barely, how the account *might* go. And that, I submit, is a great deal more than we can imagine about any alternative, noncognitive account. [I find Haugeland's discussion of the supposed alternative to cognitivism, holography, very puzzling. Holographic representations may well not represent parts of a whole by parts of themselves; but they had better be able to represent parts of a whole *somehow*. If, for example, a hologram is going to be able to deal with chess, and specifically with "moves which are powerful or dangerous," then it had better somehow encode *those very moves*, and it had better encode them as *powerful* or as *dangerous*. Thought, after all, does not operate directly upon the world, but upon encodings (or representations) of the world; that is what makes us, as the philosophers say, "intensional." I fail to see how holography yet presents a real alternative to cognitivism as a means of operating rationally on these representations. Perhaps what Haugeland means to argue here is that holograms would in some way violate his Goodmanesque, "discreteness" constraints on an IBB (e.g. definition one, section three); but that should then argue more against the appropriateness of those constraints than against cognitivism. I see no reason why a cognitive theory should not concern itself with computations defined over "analog," over "imagistic," or even over "holographic" representations: it is the computations and representations that are essential to the project, not any one particular character they might have.] But, in any case, mere skepticism is not counter-argument: I fail to see here the seriousness of the hurdles cognitivism must confront, other than simply the seriousness of actually succeeding.

ACKNOWLEDGMENT

I am grateful to Ned Block for sympathetic criticisms of an earlier draft of these comments.

REFERENCE

Schachter, S., and Singer, J. E. Cognitive, Social, and Physiological Determinants of Emotional States. *Psychological Review*. 69:379–99. 1962.

by **Richard Rorty**

Department of Philosophy, Princeton University, Princeton, N.J. 08540

A middle ground between neurons and holograms. I think that Haugeland's distinctions help us to see more clearly what it would take for "cognitive psychology" – construed as "the information processing hypothesis" – to succeed. The claim that pattern-recognition achieved through something like reflection of light between holograms would be a case in which "there are no distinct functional components whose relevant interactions are confined to intentionally interpreted articulated typologies" seems to me true and important. Haugeland has shown that the "What else could it be?" challenge may have sensible answers.

However, I have some doubts about section 7 of his paper. This section claims that moods, skills, and understanding are "potentially serious hurdles" for cognitive psychology. These do not seem hurdles of the same sort. I agree that moods can be thought of neither on the analogue of felt qualities (which, as Haugeland says, can be thought of as inputs to the cognitive IPS) nor as somehow being cognitive. But I should think they could be "incorporated in a Cognitivist explanation" by treating them as analogues to variations in the power supply to a computer (too strong or too weak voltage, or some more complex malfunction). Since moods can easily be produced by drugs, it seems fairly easy to think of their effect on cognition as being like the effect of power variations on the efficiency of a computer. There is no reason why cognitivism should think of anything that affects ability to process information as itself the processing of information.

When it comes to skills, I think that Haugeland should distinguish more sharply between, say, the professional typist and the chess master. The former's skill seems to be one that does not require cognitivist explanation, because it is so easy to think of it in simple reflex terms (connections between retinal images and finger motions, or something of the sort). For such skills, physiology or behaviorism seems enough. Cognitivism would be superflous. I find it fairly easy to think of my own ability to touch-type as a matter of rewiring (and thus of what Haugeland calls "de-interpretation" of myself as IBB) rather than as a matter of information-processing. I doubt that turning *these* sorts of skills back over to the physiologist or the behaviorist would "narrow dramatically the scope and interest" of cognitive psychology.

For skills at chess, conversation, or literary composition, however, I think Haugeland is right. These skills, however, can be coalesced with what he calls "understanding" or "insight." The common element is the ability to do something unexpected and successful. Such skills involve what Aristotle called "mastery of metaphor," and the notion of a computer program for choosing metaphors does seem crazy. It seems crazy for the reason Haugeland spells out concerning "a rationale for the insightful outputs." Here I suspect nobody really believes that the IPS model will be of much use (whereas something like holograms seems much more promising). So I think that the question that Haugeland poses might best be put as: Is there a middle ground between what can be turned over to the physiologist (for example, moods) or the behaviorist (for example, routine skills) and what must wait upon the discovery of something like holograms in the brain (for example, "brilliance")? If so, is this middle ground broad enough to make cognitivism a worthwhile program? These seem to me very good questions indeed, and we should be grateful to Haugeland for getting them into focus.

by **Robert Schwartz**

Department of Philosophy, Brooklyn College, City University of New York, Brooklyn, N.Y. 11210

Some limits and problems of cognitivism. If cognitivism were merely the position that psychology should not be limited to describing input-output relations, and that it is a legitimate scientific enterprise to concern oneself with internal states and processes, the status of cognitivism would not be very much in question. After all, even physics cannot get by without postulating internal theoretical states. But I think Haugeland is correct that cognitivism, or at least one popular version of it, typically involves the additional claim that the internal states it postulates are intentional states and the processes it theorizes about involve the transfer and manipulation of information. The cognitivist's claim seems to be that psychological explanation should involve internal states that have content or "meaning" in the way that words or beliefs are thought to have semantical or representational properties. On Haugeland's account, cognitive explanations are explanations via quasi-linguistic representations, or as Fodor (1975) puts it, cognitivism is a theory of propositional attitudes requiring a representational medium or "language" of thought.

Now if all that this fuller claim amounted to was that human beings have beliefs, desires, plans, and expectations, that is, propositional attitudes, and that ascription of propositional attitudes is easiest and clearest where linguistic behavior can be brought in, the cognitivist's position would not be very novel or that controversial. After all, appealing to propositional attitudes to describe and explain human behavior predates computer technology by centures, while drawing arrows between boxes labeled "storage," "sensory input," and "speech control" adds little to such commonplace remarks as that which we say is a function of what we see and remember. Furthermore, although a correct analysis of propositional attitudes is not at hand, few would deny the practical usefulness of talking about human activity in these terms.

Where I think cognitivism becomes a more controversial thesis is in its own analysis of intentional states and in its attempt to explain psychological phenomena other than propositional attitudes in terms of informational states and processes. Haugeland points out in the first section of his paper a few of the difficulties involved in assigning representational status to internal states and raises doubts about the scope of cognitive-type explanations in the second. While I have qualms about aspects of Haugeland's claims, my purpose in what follows is not to criticize Haugeland but to develop his points a bit more (perhaps further than Haugeland himself would feel comfortable).

To take Haugeland's second point first, if cognitive explanations are essentially explanations by means of intentional states and informational processes, they will be limited to those phenomena that can be explained within such a framework. But it is not obvious how much of what cognitivists are tempted to theorize about falls within this domain. Haugeland indicates some areas he feels may not be amenable to this sort of account: moods, skills, and understanding. Now if Haugeland's intuitions are right, this would be a damaging blow, since so many of the phenomena cognitivists hope to explain can reasonably be considered skills (e.g., pattern recognition, chess mastery, perhaps even language competence). Yet Haugeland's skepticism is not without foundations, for on the face of it, the assumption that skills or their acquisition are in general reducible to quasi-linguistic representational states is highly problematic.

While Ryle (1949) has taken a beating over the years for his atheoretical approach to explanation, the main thrust of his challenge to what he called "the intellectualist legend" still has its merit. For a wide range of skills, attempting to explain "knowing how" (i.e., a skill) or the acquisition of know-how in terms of "knowing that" (i.e., informational states) seems a nonstarter. The point has been made over and over again that knowing the principles of balance or being able to write out a grammar for L is neither necessary nor sufficient to confer upon the possessor bicycle-riding or language competence. Nor does the logic of the situation change if the knowledge is not conscious but is instead a description in some quasi-linguistic brain language. If conscious "knowing that" is not enough to account for skills, unconscious "knowing that" would not seem to be on any better ground. Possessing a skill is just not the same sort of thing as possessing information describing skill performances. What is missing, of course, is an account of how one applies the information or puts it to use. Nor is it clear in many cases what is gained in the way of explanation by postulating these internal informational states. Consider, for example, pattern recognition skill. We want to understand how a person is able to apply a label or respond differentially to all triangles or letter A's. Now to be told that the person has an internal label or quasi-linguistic representation of these concepts still leaves open the question of how the person learns and is able to apply these internal representations. So the original problem looms again, only this time it is skill with internal representations that is at issue.

Now the cognitivist's position may be read not as claiming that skills are reducible to "knowing that," but rather that several of the processes from

input to output can be described in informational terms. Thus even a clearly physical skill such as bicycle-riding can be described in intentional language (e.g., balance sensors "determine" the degree of tilt, and, "employing a rule" of balance, send "messages" to the appropriate muscles to compensate). And what is to prevent us from describing other skills and skill-acquisition similarly? Perhaps nothing. But then nothing can prevent us from describing the knee jerk reflex in this manner, too (e.g., the sensor in the joint "determines" whether the pressure is greater than n and, when it does, sends a "message" to the effector network to move the leg). And after all, the electric eye dollar-changer might also be described in propositional attitude idiom. If there is a problem with these descriptions, it is not so much that they are false but that they are somewhat misleading. They make it seem as though the knee joint and dollar-changer have internal states that are seriously like the states we are in when we have beliefs or take propositional attitudes towards some condition. When we claim a person represents something as P or believes that P, we assume the person must understand the label or proposition P. On the other hand, it does not seem reasonable to claim that these mechanisms *understand* what it means for a pressure to be greater than n or for something to be a dollar.

It is at this point, too, that one can perhaps question the force or significance of the essential cognitivist claim that psychological explanations are in terms of intentional states. That intentional idioms can be used in some descriptions of skills and mental processes need not be denied. What seems less clear is that all the internal states typically invoked have reference content or "meaning" and can be said to be *understood* in the same way that persons are said to understand words and use them to formulate beliefs.

While this is no place to launch an analysis of intentionality or understanding, I think Haugeland's demand that input-output relations "make sense" may be took weak. As he himself recognizes, almost anything, even a flipped coin, can be *interpreted* in a way that makes sense to *us*. Haugeland says such cases can be ignored, and I agree. They seem to lack the complexity, the means-ends adjustment, the features of adaptation and learning, of being subject to reason and evidence that we see as part and parcel of our more usual attributions of intentionality and understanding. To give substance to the claim that internal states are intentional, it must be reasonable to see these states as having sense and making sense to the organism or mechanism using them, not just to us. The fact that a cell complex fires whenever condition ϕ occurs is not to claim that the complex understands ϕ. Nor is the situation substantially changed if the complex not only fires but produces something else that because of the correlation, we are tempted to call a ϕ label or representation. For the ϕ label will not be very intentional unless it is itself understood. Furthermore, as these internal states become more removed from correlation with external stimuli, it becomes increasingly difficult to see them as understood representations having reference and other semantical properties on par with our more ordinary symbols.

Understanding is not an all-or-none process or the grasping of some fixed idea. Understanding, to borrow a phrase from Harman, involves mastery of an "evidence-inference-action game" and is itself very much a skill-like affair. Now to the extent the use, function, and acquisition of our internal quasi-linguistic representations do not share these features, the claim that these representations are understood seems problematic. In turn, the claim that our internal states and processes are truly intentional rests on the analogy between these states and our more usual propositional attitudes. At the point where the analogy breaks down, we can begin to question what remains significantly distinctive to cogntivist explanations.

REFERENCES

Fodor, J. A. *The Language of Thought,* Thomas Y. Crowell, New York, 1975.

Harman, Gilbert. "Language, thought, and communication" in Keith Gunderson, ed., *Language, Mind, and Knowledge,* University of Minnesota Press, Minneapolis, 1975.

Ryle, Gilbert. *The Concept of Mind,* Chapter 2, Barnes and Noble, New York, 1949.

by **Thomas W. Simon**

Department of Philosophy, University of Florida, Gainesville, Fla. 32611

On Cognitivism's explanations and limitations. Haugeland's task is twofold: (1) to provide an explication of cognitive theory and research, and (2) to appraise (largely negatively) the cognitivist research program. Although most of Haugeland's exposition of cognitivism appears correct, some of his more critical attempts are problematic, particularly the alleged "potentially serious hurdles" to cognitivism. Before turning to these, a few emendations and criticisms of Haugeland's treatment of reductionism are in order.

While agreeing with Haugeland that "the ills of psychology are [sometimes] laid to a misguided effort to emulate physics and chemistry," the importance of physics and chemistry to cognitive psychology should not be underplayed. Making these connections helps undermine the frustration often associated with the black-box nature of theorizing in cognitive psychology. How, for example, does one choose between two radically divergent interpretations of the component Intentional Black Boxes (IBBs)? The physical sciences can provide a strong candidate standard in terms of which that choice can justifiably be made.

Accordingly, Haugeland's three explanatory maneuvers can be ranked in terms of their explanatory power. Explaining an IBB as an Information Processing System (IPS) occupies the lowest rung, for such an account is limited to the particular kind of task in question. Next is explaining an IBB by intentional instantiation, which is the level of explanation claimed for the General Problem Solver computer simulation. These two levels, however, exemplify what Cummins (1977, p. 272) calls the Analytical Strategy of analyzing a disposition into a number of relatively less problematic dispositions such that organized manifestations of these analyzing dispositions amount to a manifestation of the analyzed disposition. Moreover, subsuming dispositional regularity under a physical law (the Subsumption Strategy in the Cummins's terminology and explanation by physical instantiation in Haugeland's) provides the third level of explanation and constitutes an important, if not critical, constraint on the first two levels (the Analytical Strategies). Here, it is worthwhile to quote Cummins (p. 273) in full:

"A natural assumption – and a correct one I think – is that the Analytical Strategy must eventually terminate in dispositions which yield the Subsumption Strategy. For without this assumption, the apparent explanatory progress afforded by the Analytical Strategy is *mere* appearance. That strategy makes progress only insofar as the analyzing capacities are relatively less problematic as compared to the capacity analyzed. We undermine such progress if we suppose that our analyzing capacities might ultimately prove resistant to the Subsumption Strategy, for to suppose this is to allow that the capacities may be utterly mysterious and inexplicable from the point of view of physical science: we shall be barred from any account of why some things, and not others obey the associated law. One needn't endorse any starry-eyed claims about the unity of science to find this prospect unwelcome."

So, contrary to Haugeland, cognitivists should be interested in psychophysical bridge laws and very worried if none are possible.

Turning from Haugeland's exposition to his critique, he proposes three "potentially serious hurdles" to a cognitive research program: moods, skills, and understanding. Wisely avoiding the pitfalls of impossibility arguments ("your research cannot, in principle, accomplish x"), Haugeland shows the implausibility of these three phenomena being given a cognitive treatment. Yet, even though an account of each will undoubtedly prove difficult and complicated, contrary to Haugeland, no prima facie reasons exist for doubting the applicability of the cognitivist approach to them.

Haugeland's case against the cognitive account of moods can only be made by assuming that moods "don't seem at all cognitive themselves." Yet, there is just as much prima facie evidence to believe that a mood is not an integral quality but is rather a complex interplay of feelings, beliefs, evaluations, and the like. While admittedly "moods permeate and affect all kinds of cognitive states and processes," what is being affective in that case may well include some other cognitive state and process in addition to some felt qualities. In other words, my mood affecting my present thinking may be tantamount to some general belief that I have about the world (for example, extreme pessimism) affecting my current thinking. There do not seem to be any prima facie reasons for accepting Haugeland's account over this one. Moreover, the extent Haugeland allows for segregating felt qualities from the cognitivist's concerns is the extent to which they can be segregated as factors of moods.

Skills present a somewhat different problem in that many skills, particularly motor skills, do not seem analyzable into discrete mental (or information) processes. Yet, this may only indicate a different kind of information processing, as, for example, pattern recognition, taking place while one is

acquiring a skill. As Simon notes (1976, p. 80), these different types of information processing account for some of the different abilities of, for example, duffer and master chess players: ". . . the short-term memory of the chess master has the same capacity, measured in chunks, as the short-term memory of the duffer, but [the] duffer's chunks consist of individual pieces while the master's chunks consist of configurations of pieces." Hence, in applying this to Haugeland's example, the duffer's and master's "consciously considering on the order of a hundred plays in thinking out a move" may simply be different ways of processing the information, and the master's ways are as amenable to a cognitive account as the duffer's.

Finally, "understanding" is highly ambiguous in that it can be contrasted with ignorance, mere knowing, and misunderstanding (Scriven, 1972, p. 32); Haugeland's sense of understanding as insight is no exception. Insight could consist in "believing or feeling that one understands" (Scriven, 1972, p. 32), in the ability to specify why something makes sense, or in developing a novel concept. None of these is beyond the purview of cognitivism. While Haugeland wants to deny that "insight is itself some 'transcendental' or impenetrable mystery, which we are forever barred from explaining," he still wants to "admit that the phenomenon of insight is simply mysterious and unexplainable at present." Some of the mystery surrounding the concept insight is dispelled by providing an analysis, and an analysis would spell out just those antecedent conditions thought by Haugeland to be implausible. Admitting that all new developments would have to be products of antecedent general conditions does not "preclude any radically new ways of understanding things." On the contrary, it helps us understand creative insights. Even Galileo's derivational-nomological explanation had important precursors (Shapere, 1974).

The enigmatic phenomena cited by Haugeland are just a sample of those that people have proposed as obstacles to cognitivism. Some general comments, negative and positive, can be made about this strategy. First of all, the strategy assumes more than it seems capable of delivering, for something fairly definitive must be known about the obstacle phenomenon that makes it incompatible with the cognitive research program. For example, claiming that insight presents a serious hurdle to simulation presupposes a fairly well-worked out analysis of the constituents of insight. After all, we want to know what it is about insight that creates the problem. Here a dilemma arises. For once that analysis is given, then the phenomenon, in being more manageable and less mysterious, seems more amenable to a cognitive treatment. Nevertheless, Haugeland's warning that the simulation approach may well be, like behaviorism, overstepping its fruitful bounds of inquiry is well worth noting. Yet, in many cases these limitations are uncovered only by overstepping them.

Although I have been mostly critical of Haugeland's paper, that should not detract from the many fruitful insights he has provided, particularly those on explanation and intentions.

REFERENCES

Cummins, R. Programs in the Explanation of Behavior. *Philosophy of Science.* 44:269–87. 1977.

Scriven, M. The Concept of Comprehension. In: R. Freedle and J. Carroll (eds.), *Language Comprehension and the Acquisition of Knowledge.* New York: John Wiley & Sons, 1972.

Shapere, D. *Galileo.* Chicago: University of Chicago Press, 1974.

Simon, H. Identifying Basic Abilities Underlying Intelligent Performance of Complex Tasks. In L. B. Resnick (ed.), *The Nature of Intelligence.* Hillsdale, N.J.: Erlbaum, 1976.

by **Charles Taylor**

All Souls College, Oxford University, Oxford OX1 4AL, England

Indivisible performances, implicit grasp, and the problem of meaningfulness. John Haugeland's excellent article has done a great deal to clarify the issue about cognitivism. In particular, his distinction within reductive explanations of different kinds of "instantiation," as he calls it, makes the process involved much clearer.

I agree very much with Haugeland that "the eventual fate of cognitive psychology will be settled empirically"; and I share with him some doubts about whether it will in the end pan out. I would like to expand a little on these. My remarks here touch on an issue that is latent in each of the three areas where Haugeland sees "potentially serious hurdles" for cognitivism in his section 7.

The great strength of cognitivism over behaviorism, as Haugeland points out, is that it promises to give acceptable scientific accounts of our rational behavior qua rational behavior. An important phase of the explanation involves breaking down our rational performances into the component steps that make them up (or, it is believed, must make them up); for instance, the steps of calculation that enable us to arrive at a correct answer. The whole performance can be explained by a series of part-performances. And there is the additional payoff, from the standpoint of scientific reductionism, that these part-performances can eventually be matched by machines, thus promising what Haugeland calls a physical instantiation.

A central difficulty for the cognitivist program that seems to be emerging more and more clearly is the existence of performances that cannot be so broken down. An example is the chess player's preselection of certain possible moves as the ones worth working out in all their consequences. It seems implausible to hold that this results from an unconscious step-by-step calculation of the chances of a much wider range of moves, or a step-by-step application of some heuristic procedure.

But nor can such immediate judgments or unconscious selections be just seen as a species of immediate experience, like our seeing red, or feeling cold. For they are cognitive performances and have their own kind of rationality. Indeed, they are related to the step-by-step particulate observations that do not seem to explain them; and that (sometimes) in two ways. First, we may build up over time an ability to grasp at a glance that this area is one of weakness for our opponent and therefore worth considering. But, typically, we start off in a more painful step-by-step way. We learn by particular experiences – our queen is pinned in one game, queen and king are forked in another – to see a certain configuration as one of weakness, even though this weakness-judgement cannot be exhaustively specified in any such list of particular disasters. The step from particular discovery to general grasp is the same we observe with many skills, as Haugeland points out, where the stage of learning may require a lot of concentration on particular movements, which concentration is not only unnecessary but impossible when we have mastered the skill.

Secondly, our grasp of the board is related to particulate judgments also in that we can make it yield some. If asked why this position looks weak, we can give a number of possible concrete dangers that could arise. And the chess player does something like this when he begins to zero in and consider particular moves. And yet, while related in both these ways to particular judgments about the consequences of particular moves, the chess player's grasp of the board does not seem to be rationalized by them, in the sense that the successful performance of this general grasp of the position results from a series of part-performances that are particular judgments.

In other words, we may have to allow for a particular kind of human performance, which is the achieving of an implicit grasp of a domain. To have such a grasp is to know something of the kind of explicitly statable fact that is true of the domain, but not necessarily to have (even unconsciously) registered any such facts. To put it another way, there is a relation of justification between any claim to an implicit grasp and explicit factual judgments – for instance, my claim to grasp areas of strength and weakness on the board is considerably less credible than that of a master, just because my hunches so often turn out badly when I actually work through the detailed moves. But this relation of justification is not matched by an explanatory one – I do not have the grasp *because* I have made the explicit judgments.

The implicit grasp seems to play an important role in human life: think of all the things we are always taking in and responding to about the feelings of our interlocutor, the mood of a conversation, the beauty of a landscape and a work of art, where any attempt to be explicit seems an attempt to unpack very partially our implicit understanding, rather than to provide the grounds that determined the judgment. If such performances turn out to be something like what they look or feel to us – and the pattern of failures of Artificial Intelligence and Computer Simulation make this more plausible – then they will not be amenable to the mode of explanation hitherto dominant in cognitive psychology. The issue is not mainly one of atomism versus holism. It is not because these performances have a gestalt property, and are undivided wholes, that they pose a difficulty. The problem is rather that they seem to have their own kind of rationality. They *are* related to particulate judgments, that is just the trouble, only not in the canonically prescribed way.

If these performances do turn out to be unamenable to the cognitivist approach, then this will also involve our going beyond the account of meaningfulness that Haugeland gives in section 3. This in two ways. First, the clear

singling out of a range of tokens that can then be assigned to a number of types may in the end fit only the linguistic cases, or at least the cases where we have some notation. But when we are, for instance, grasping someone's mood, or his feeling toward us, in his speech and manner, we find it very difficult, if not impossible, to reconstruct the performance in two stages: the selection of the cues, on one hand, and the interpretation given them, on the other. And this may correspond to a basic and not just surface feature of this kind of performance.

The second possible modification concerns what it means to make sense of a given domain through an interpretation. Haugeland talks of making reasonable sense of outputs in the context of prior inputs and outputs, and sees the capacity to do this as the test of an interpretation. The example here is a chess game. By interpreting "P-K4; P-K4; QKt-B3 . . ." as a series of chess moves, we can see how all these make the kind of sense that a chess game makes. The difficulty arises around the kind of sense that will be recognized as such by cognitivist psychologists. They can certainly recognize various overall patterns as making sense, that is, satisfying specified cogency conditions, and this can be the meaning they give to "meaningfulness," provided the elements forming the pattern are not themselves defined in terms of intentionality or meaningfulness. Thus the moves of a game of chess form a pattern, defined quite independently of what it is like to experience or be aware of a given event as a chess move. Indeed, when machines play chess we do not usually imagine that anyone is experiencing or being aware of the moves. Similarly, the theories of meaning developed out of the work of Quine and, latterly, Davidson, are grounded in the coherence we can see in a set of utterances in a given context, where the coherence is understood as holding between the truth-conditions of the utterances and the facts that obtain. Once again, this is a pattern we can discern in things that are defined quite independently of what it is to understand, or be aware of, the meaning of an utterance. The Quine-Davidson type theory of meaning is very much one that might be developed by an outside observer who never entered into communication with the beings whose language he is trying to map.

It is doubtful whether this notion of meaningfulness could cope with interchanges founded on the kind of performance I called an implicit grasp, for instance a conversation where I somehow take in and sustain the tone my interlocutor has set. For perhaps the tone itself can be defined only in terms that refer to meaningfulness. This latter could then no longer be defined in terms of a certain cogency without circularity. This is not meant, of course, as a criticism of Haugeland's careful and illuminating discussion in section 3. On the contrary, he presents there very well the kind of notion of meaningfulness that cognitivism requires. My doubts concern whether this notion is ultimately defensible; and for all I know, Haugeland may share these doubts.

The question of principle raised here is whether we can have an adequate theory of meaning that does not give an account of what it is to understand an expression, or be aware of something as meaningful. This issue has already been raised against Davidson's theory of linguistic meaning, for instance by Dummett (1975). There may be another key issue here that cognitive psychology will have to face.

REFERENCE

Dummett, M. What is a theory of meaning? In: S. Gutenplan (ed.), *Mind and Language*. Oxford, 1975.

by **Ryan D. Tweney**

Department of Psychology, Bowling Green State University, Bowling Green, Ohio 43403

Is making reasonable sense reasonable? It should be said at the outset that Haugeland's fine analysis of cognitivism does not encompass all varieties of what is ordinarily called "cognitive psychology." That is no criticism, but it is important to see the limits of application of his view. Thus, a good part of contemporary research *is* aimed at what Haugeland calls Deductive-Nomological Explanation. Consider, for example, chronometric experimentation on mental processes, a tradition that Blumenthal (1977) had characterized as deriving from Wundt. Further, Haugeland's approach takes no account of those explanations based on developmental processes. Piaget's (1970; see also Brainerd, this issue), in particular, does not fit comfortably with any of Haugeland's three types of explanation, since none are time-dependent. This having been said, it is important to see that his analysis does relate, in an interesting fashion, to all those attempts made in recent years to provide systematic explanations of cognitive phenomena. It thus covers simulation-dependent approaches like Simon's (1969 *op. cit.*), semantic-based models of memory like Norman and Rumelhart's (1975), and much else in addition.

Haugeland attempts to characterize such cognitive psychology in a way that is "unimpeachably rigorous and empirical all the same." The heart of his argument is the use of intentional interpretation as an explicit formalism permitting the incorporation of meaningfulness in cognitive models. Put briefly, if an intentional interpretation of an ordered behavioral system is possible, and if the interpretation "makes reasonable sense," then we can use the interpretation as part of a systematic explanation. Are such interpretations testable? If they are, that is, if they lead to potentially falsifiable claims that can be checked, then cognitivism will, indeed, rest on an empirical footing. Yet, how can we accept "making reasonable sense" as a criterion? Haugeland feels that this presents no difficulties: "By and large everyone can agree on what does and doesn't make sense."

This is, however, simply not so. To see the problem, consider the human propensity for detecting meaning and pattern where none exists. This is a propensity that forms the basis of all projective tests. It is, in fact, extraordinarily difficult *not* to see a pattern. If you doubt this, try looking at a tiled floor without grouping separate tiles into small groups, or try seeing a Rorschach card as "just an ink blot." These elementary phenomena are part of a broader tendency, one that means we will always be able to come up with an interpretation that "makes reasonable sense." Further, social agreement on which interpretations make the most sense is no help at all. Consider the Necker cube. By Haugeland's criterion, it is truly a *three*-dimensional object, since most people see it that way.

A similar problem has confronted psychology before. In 1889, Binet wrote that all animals, even one-celled protozoa, possessed psyches. Since such creatures show intentionality in pursuit of food, in sexual interaction, and in avoiding noxious stimuli, it "made reasonable sense" to Binet to attribute the same psychological structures to protozoa that are found in man. Was Binet *wrong?* We would now say yes, but not because his notion of mind was testable (it was not) and not because it did not make reasonable sense (it did). It failed, instead, on grounds of parsimony. There is an alternative approach – Morgan's (1891) in particular – which offered a better way to understand protozoa. Even if we disagree that Morgan's Canon applies to human thought, the notion of parsimony continues to compel assent. If, as Haugeland says, a nonintentional model of mind can be developed, it will create grave problems for cognitivism, and for exactly the same reason.

The problem with cognitivism is not that patterns of behavior do not exist. The existence of chess games and of, say, de Groot's (1965 *op. cit.*) finding of memory differences between master players and amateurs is sufficient evidence to refute elementist approaches to cognitive phenomena. There *are* patterns, and the patterns do call for explanations. What I am arguing is that intentional explanations of the sort developed by Haugeland have not been shown to be testable in the way he has argued. Intentionality does, indeed, "make reasonable sense," but we cannot admit *that* as our touchstone for testability. Descartes' "Evil Genius" lives within us and will fool us every time if we try.

Haugeland has not, of course, simply presented a defense of cognitivism. The most valuable part of his presentation concerns three potential empirical hurdles that cognitivism must overcome: moods, skills, and understanding. In calling these the "tips of some of the icebergs on which (cognitivism may) founder," Haugeland may be alluding to the common problematic characteristic of all three problems, namely consciousness. We still lack any real sense of the role that consciousness should play in thinking. None of our models, least of all computer simulations, provides a functional role for consciousness. It is this gap in our approach to cognitive phenomena that provided a beginning point for Jaynes's (1977) startling book on consciousness as something to be explained. But how? Deductive-Nomological Explanations? Systematic Explanations? Or Jaynes's unique Phylogenetic Explanation? Haugeland has helped us see that sooner or later the problem will need to be directly addressed by cognitive psychology. We may never be able to say what consciousness *is*, anymore than biologists have been able to say what life is. But we should be asking, more often that we do, what attributes consciousness possesses and what relationships it enters into. Perhaps then we will be able to look forward to not our Newton, but to our Watson and Crick.

REFERENCES

Binet, A. *The Psychic Life of Micro-Organisms: A Study in Experimental Psychology* (T. McCormack trans.) Chicago: Open Court, 1889.

Blumenthal, A. L. *The Process of Cognition.* Englewood Cliffs, N.J.: Prentice-Hall, 1977.

Jaynes, J. *The Origin of Consciousness in the Breakdown of the Bicameral Mind.* Boston: Houghton Mifflin, 1977.

Morgan, C. L. *Animal Life and Intelligence.* Boston: Ginn & Company, 1891.

Norman, D. A., and Rumelhart, D. E. *Explorations in Cognition.* San Francisco: W. H. Freeman, 1975.

Piaget, J. *Genetic Epistemology* (E. Duckworth, trans.). New York: Columbia University Press, 1970.

by **Ernst von Glasersfeld**

Department of Psychology, University of Georgia, Athens, Georgia 30602

Some problems of intentionality. Haugeland's paper will be welcomed by seasoned as well as incipient cognitivists as an eminently helpful and stimulating contribution. If it should not convert staunch followers of the behaviorist gospel to the view that "the Cognitive approach to psychology offers . . . a science of a distinctive form," it will probably not be Haugeland's fault. His lucid exposition of the three types of explanation cuts through much of the fog that has been created by enthusiastic but often inaccurate promulgation of cybernetics, systems theory, and structuralism – all of which imply the discrimination between "morphological" and "systematic" coordination. Haugeland should also be congratulated on his unequivocal statement of the fact that explanatory reductions do *not* "supplant the explanations they reduce" and on his courageous frontal approach to the problem of intention.

Using chess as an example in his analysis of how one may come to interpret a black box as intentional is a good didactic simplification. I would suggest, however, that the very element whose exclusion makes that example simpler than other, less conventional activities might lead us to take a somewhat less positivistic stance than does the author in his later evaluation of Cognitivist theory.

Condition (iii) for interpreting an object as an IBB is that the object's outputs "consistently make reasonable sense." Haugeland is aware of the problems this expression raises, but he says that "it is seldom hard to recognize in practice" how "reasonable sense" should be defined. I think we have to be more explicit about this. In the context of several games of chess it will, as a rule, be easy to decide whether or not a presumed player's moves, on the whole, make sense. There will even be occasions when this can be decided about a single move. This is so because within the context of chess we *know* a priori what a player's goal has to be, and there will be no doubt whatsoever about recognizing it when it is achieved. It is, indeed, a matter of accepted *rules,* and a person or box that has no conception or a deviant conception of what constitutes "mate" will not be considered a chess player at all. However, when we come to consider other activities that are not so obviously governed by a set of explicit conventional rules, the situation is much more obscure because we have no a priori knowledge of the observed subject's goals in terms of which his or her actions could be judged to make sense and, hence, to be intentional.

Jurists, who are frequently faced with the problem of deciding whether or not a person's action was intentional, have created the rather powerful maxim: A person will be presumed to intend the natural, probable consequences of his acts. This works well enough in court, because there it is tacitly assumed that people have much the same ideas (knowledge) as to what are natural, probable consequences of the acts under consideration. But if, as philosophers or scientists we are faced with a black box, a wild-living chimpanzee, or a person from a significantly different cultural background, we are in no way justified in making that tacit assumption, because we simply do not know what they *believe* to be natural, probable consequences of acts. Therefore, as long as we remain passive observers, we cannot be sure whether or not their acts are intentional (cf. von Glasersfeld & Silverman, 1976).

Fortunately, as Hofstadter (1941) has suggested, there are ways and means for an observer to test hypotheses about an observed organism's intentions or goals, by creating obstacles or, more generally, disturbances for the organism (cf. Powers, 1973, for tentative implementations). Epistemologically, such tests are equivalent to any other hypothesis tests, in that they may tell us whether or not our hypothesis remain a *viable* explanation for our observations, but not whether this explanation is "true" or "false" in any absolute ontological sense.

There is yet another epistemological aspect to be considered. The intentions that an observer attributes to an observed organism are necessarily determined in a limiting sense by the set of goals that are *conceivable to the observer* as well as by *his* beliefs concerning rules, methods, and activities that are likely to lead to the attachment of these goals. This observer-dependent "conjectural" character of explanations, however, can hardly be said to be unique or discrediting to the Cognitivist approach to "systematic explanation." It seems to be the character of all scientific explanation.

Another fairly important issue is involved in the hypothetical example of the brain hologram. The proposed "fixed" association between an "important substructure in chess positions" and "powerful or dangerous" moves would have to be considered either innate or acquired. If we decided that it was innate, we would have to discard any intentional explanation because, as long as we believe the theory of evolution, we would be obliged to say that the association was the result of accidental variation and unintentional selection. If it was acquired, however, it would have to have been "holographed" into that brain at some prior time by someone's goal-directed, intentional selection. If the brain now shows no evidence of making *new* selective associations of that kind, it would be classified as an IBB whose goals have been set by some IPS that decided which positions were to be associated with which moves. If, on the other hand, the brain were still making new selective associations, the information-processing and decision-making capability would have to reside within it (at least if we exclude hypnosis or telepathy) and, qua whole brain, it would be classified as IPS. This distinction, I hasten to add, is obviously based on the distinction Pask (1969, p. 23) made between "purpose of" and "purpose for."

That point is relevant also to the problem Haugeland raises with regard to skills. I can see no reason why an IPS of a certain level of complexity should not have the capability of "deliberately and thoughtfully" building up, say, a sequence of motor acts under appropriate perceptual feedback control for a given recurrent purpose, and then "automating" the whole arrangement by giving it direct access to the sensory signals that were originally used for the control of the activity in the central processor. Since we can build a thermostat that perfectly realizes the *purpose of* maintaining, without consciousness on its or on our part, the room temperature we set (purpose for), there seems to be no *logical* obstacle to our automating (after deliberate compilation) the general motor pattern for hitting a ping-pong ball in such a way that only those parameters that determine *where* the ball with go remain under our conscious control. People who have to learn to double-declutch when shifting gears (e.g., in competitive sports car racing) seem to do exactly that. The movements of foot on clutch and hand on gear lever come relatively quickly under autonomous control; gauging the intermediary jab on the gas pedal according to perceptual signals indicating engine rotation and actual speed of the car, however, takes very much longer to become "automatic" and probably never does so entirely. The salient feature in all this is that, experientially, we are all aware of the fact that there are many quite complicated motor activities whose control, after a period of more or less conscious supervision, can be relegated to an unconscious level. The application of control theory to these cases seems promising because it supplies a conceptual model for the strange observation that, while the execution of the motor sequence seems wholly unconscious, their direction (i.e., the setting of particular goals of reference values that determine each individual execution) remains under conscious control.

It is to be hoped that Haugeland's paper, because it supplies a number of very clear methodological definitions and draws attention to problems that psychologists, by and large, would rather avoid, will not only be discussed but also acted upon. There are, I believe, good reasons to predict that both empirical and theoretical investigation of the "hurdles" he mentions will show the Cognitivist's systemic approach rather more powerful and fertile than this conservative evaluation might lead one to expect.

REFERENCES

Hofstadter, A. Objective teleology. *Journal of Philosophy.* 38(2):29–39. 1941.

Pask, G. The meaning of cybernetics in the behavioural sciences. In: J. Rose (ed.), *Progress of Cybernetics.* Pp. 15–44. New York, Gordon and Breach, 1969.

Powers, W. T. *Behavior: The Control of Perception.* Chicago, Aldine, 1973.

von Glasersfeld, E., and Silverman, P. Man-machine understanding. *Communications of the Association for Computer Machinery (Forum)*, 19(10):586–87. 1976.

by **Catherine Wilson**

Department of Philosophy, Barnard College, New York, N.Y. 10027

Cognitivism's contributions: some questions. The most clearly focused sections of Haugeland's paper provide a potentially valuable account of the disunity of science. His main thesis is that "cognitivism," understood as the view that intelligent behavior can be explained in terms of "internal cognitive processes," has suggested an entirely novel approach to the question of explanation in psychology. ("Intelligent behavior" has here to be construed broadly enough to include such different activities as understanding language, playing games of strategy, and visually recognizing objects.) If I have understood it correctly, his view is that cognitivism has liberated us from the paralyzing assumption that explanation in psychology must be "derivational-nomological." Instead of looking for laws and axioms governing behavior, we are entitled to look for explanations of "how things work," given in terms of the interaction of discrete functional components, that is, for "systematic" explanations.

My first problem is a direct result of the author's ambivalent assessment of the value of computer-simulation studies. On one hand, he gives an interesting argument to the effect that what is involved in being in a certain mood, or having a certain skill, or acquiring insight into a problem cannot be explained as a feature of an information-processing system. This in turn casts doubt on the assumption that even routine problem solving can be explained as the activity of an IPS, as long as moods, skills, and insight can be expected to affect directly the ability to solve routine problems. It would be tempting to conclude that such simulation studies lack theoretical value for psychology, and it is surprising to find the author maintaining just the opposite, in particular that such studies have greatly contributed to our understanding of the problematic notions of "meaning" and "significance." It is difficult to tell from the exposition how the improved notion of meaning differs from the behaviorist notion of "stimulus-meaning," while remaining powerful and empirical. Perhaps the ability of an IPS (or even an IBB) to produce novel but interpretable outputs in response to novel inputs has an explanatory role to play here, but it would be helpful to have some fairly informal clarification.

My other difficulty concerns the possibility of what might be called "cognitivism without automativism." Here I am troubled by the possibility of arguing in the following way: "If human behavior could be explained as the behavior of an IPS, then 'systematic' explanation in psychology would be possible . . . But there is little reason to suppose that most intelligent or even goal-directed behavior does involve the representation of problems as branching tree-structures, and the performance of subsequent surveying and matching operations . . . Therefore, there is no reason at all to suppose that explanation in psychology can be systematic. What science is possible may turn out to be unified after all."

No matter how easy it is to refute such an argument, it requires refutation because it has been energetically maintained by philosophers, for example, Ryle, that systematic explanations of rational behavior are impossible. The fact that many people find Ryle's arguments somewhat weak in this connection is entirely beside the point. If the position is false, it must be false because psychology has already succeeded in showing that such apparently simple concepts as "perceiving," "understanding a sentence," and "deciding on a move" can be understood as processes that involve the cooperative interaction of a number of separate subprocesses. If there is any reason to suppose that explanation in psychology is distinctive in the way Haugeland suggests, examples should be ready to hand. It is extremely difficult to evaluate the suggestion that holographic models might provide explanations of the right kind without knowing how the model is supposed to be realized in the actual structures of the brain. In particular, it is hard to know whether the form of explanation involved will be "systematic," "morphological," or something else altogether.

To sum up, then, my main questions are two. Haugeland makes it clear that, as far as he is concerned, recent studies in artificial intelligence have more than a purely practical value; they do not simply give us a reliable means of testing, say, the adequacy of grammars: What exactly, then, are the virtues of cognitivism, which make it, in spite of its implausibility, subject to neither the sterilities of behaviorism nor the vacuities of introspectionism in its treatment of "meaning"? Second, precisely what difference does the truth or falsity of the theory that human beings can be viewed as extremely complex IPSs make to the question of explanation in psychology?

by **Miriam L. Yevick**

Department of Mathematics, University College, Rutgers The State University, Hill Center, New Brunswick, N.J. 08903

The two modes of identifying objects: descriptive and holistic for concrete objects; recursive and ostensive for abstract objects. In the section in which Haugeland proposes a hologram-like process as a possible alternative to the I.P.S. model of cognitive behaviour, he writes as follows: ". . . there is nothing going on which can be regarded a 'working out a rationale with quasi-linguistic representations.' By contrast a typical computer based pattern recognizer is an I.P.S. Thus searching for discontinuities in luminance gradients, proposing that they are edges, checking for connexity and so on [i.e. procedures usually called feature extraction, -my addition] are all rational procedures relating to the *problem of identifying objects.*"

In Section 7, he states: "First, with rare exceptions, articulateness about a *skill*, no matter how detailed nor in what specialized quasi-linguistic notation, is neither necessary nor sufficient for having it; it always takes practice and often expert examples and talent ($\neq$ intelligence)."

The author here points out a distinction between two modes of understanding our environment: the first identifies objects by quasi-linguistic representations; the other apprehends objects by means of nonarticulate skills. This dichotomy; which is undoubtedly related to the complexity of the concrete objects to be recognized or manipulated, was projected as follows by von Neumann (1966, pp. 51–54): "certain objects are such that their description is more complex than the object itself."

We can explicate this proposition on a theoretical level in the domain of optical patterns. (See Yevick, 1975 *op. cit.*). Such patterns or objects are thin, white regions on a black background. These can be simple (regular), like the outlines of rectangles; or complex, like the outlines of Chinese characters or random-like motions. The following holds true: a complex object requires a long (sequential, quasi-linguistic) description but yields a sharp recognition (auto-correlation) spot under holographic filtering; hence it is identified most readily by holographic recognition, or holistically. A simple object requires a short (quasi-linguistic) description but yields a diffuse recognition spot; hence it is identified most readily by quasi-linguistic representation or description.

Description and holographic recognition thus appear as two (complementary) modes of identifying an object: the more complex the object, the longer its description and the sharper its auto-correlation spot, and vice versa. The more complex they physiognomy of a person, the more unique, and hence sharper, its identity and ease of recall; the more simple, the more common and hence "unidentifiable." Perfect holographic recognition obtains for a totally "random object", that is, one with an infinitely long description; for a perfectly sharp point the opposite is true.

Suppose that one is given a store of objects with which one is familiar, a holographic recognition device, and a quasi-linguistic mode of representation; one is then presented with an arbitrary object to be "identified." An approximate match is obtained either by producing a description of acceptable length or by holographic recognition of a subset of similar (associated) objects from the store. The mode of identification that will be more appropriate then depends on the complexity of the unknown object. If it is simple, we "know" it by a short linguistic description; if it is complex, by the "associations" it evokes.

If we consider that both of these modes of identification enter into our mental processes, we might speculate that there is a constant movement (a shifting across boundaries) from one mode to the other: the compacting into one unit of the description of a scene, event, and so forth that has become familiar to us, and the analysis of such into its parts by description. Mastery, skill and holistic grasp of some aspect of the world are attained when this object becomes identifiable as one whole complex unit; new rational knowledge is derived when the arbitrary complex object apprehended is analytically described.

This dichotomy is also relevant in the purely formal domain where we deal with so-called abstract objects. We note that Haugeland introduces the notion of "dimension" and discusses the "disconnected" nature of two orthogonal dimensions. He then applies this insight to indicate the incom-

parability of the mind of a mathematician and the Turing-Machine, and relates this to Gödel's proof that any formal system rich enough to include arithmetic can express truths not provable in the system.

A careful scrutiny of the various presentations leading to Gödel's result reveals that the "abstract objects" that are the entities under discussion in a formal system actually occur in two modes: as objects identified by bold-faced pictures or shapes or marks on paper, and as objects generated recursively from certain zero-entities, recognized in some way by their rank, that is, the first of a certain list (Quine, 1950); an expression of length one (Shoenfield, 1967); a sequence of one symbol (Gödel, Pred. 15 in van Heijenoort, 1970; Mendelson, 1964); entities generated by a successor operation on a pair of arguments (Kleene, 1970, pp. 247, 251–252; Pred. Dn 1 should read: $y \neq 0$). We recognize abstract objects of the first kind by ostension (holistically); those of the second kind are recognized by sequential generation or description. But whereas in the case of concrete objects discussed above, it is possible to assert that these two modes of (approximate) indentification refer to the same object, the abstract objects have no identity recognizable beyond their formal mode of representation or generation. Thus, going beyond Haugeland's remark and using the word "mode" for his "dimension," we note that the mixing of modes is already present in the argument that yields Gödel's undecidability result: it rests essentially on the identification of abstract objects ("formal numerals") recognized in two disconnected modes. The well-known confusion between Mention and Use reappears here as a confusion between showing and telling or display and enumeration, that is, as a mixing of dimensions.

The following quotation from Freudenthal (1960), who attempted to construct a language, "Lincos," aimed at cosmic communication, clearly projects the irreducible duality: "We have agreed to abstain as much as possible from showing (concrete things or images of concrete things) but we cannot entirely abstain from it. Our first message will show numerals as an introduction to mathematics. Such an ostensive numeral, meaning the natural number n, consists of n peeps with regular intervals; from the context the reader will conclude that it aims at *showing* just the natural number n."

For minds to communicate or to do formal mathematics, they must possess both a quasi-linguistic (sequential, rational) and a holistic (ostensive, associative) dimension.

REFERENCES

Freudenthal, H. *Lincos*. North Holland: Amsterdam, 1960.

Kleene, S. C. *Introduction to Metamathematics*. New York: Van Nostrand, 1970.

Mendelson, E. *Introduction to Mathematical Logic*. New York: Van Nostrand, 1964. Pred. 14 b.

Quine, W. *Introduction to Mathematical Logic*. London: Norton, 1950.

Shoenfield, J. R. *Mathematical Logic*. Reading, Mass.: Addison Wesley, 1967.

van Heijenoort, J. *Frege and Gödel*. Cambridge, Mass.: University Press, 1970. Pred. 17.

von Neumann, J. *Theory of Self Reproducing Automata* (A. W. Burkes, ed.) Urbana: Univ. of Illinois Press, 1966.

Author's Response

by **John Haugeland**

The critical assessment of Cognitivism: a closer look

Like most of the commentators, I shall concentrate my remarks on specific issues where I have detailed rejoinders, emendations, or concessions. But I do want first to express my genuine thanks for the many generous and thoughtful observations made by so many people; I hope this necessary brevity does not obscure my sincerity. Since a number of questions were raised by more than one author, and still others touch on similar problems in interestingly related ways, I have chosen to organize my replies by topic, rather than by commentary; for convenience, these are arranged in roughly the order of the original paper. To facilitate scanning for particular points, commentators are cited in bold face type.

Section 1. Arbib objects to the distinction of explanatory styles because they "can occur in virtually any combination." But his example of a ball rolling on a surface depends on the characterization of the surface. A mathematical solution with the surface defined only as H(x,y) would be derivational-nomological (D-N), whereas an explanation of a path that cited only the walls of a tilting maze, or the grooves in a model drainage system, would be morphological; and there could be intermediates. The analysis of DNA "involves physics as much as ball-and-stick chemistry" because it involves not only an explanation of the replication ability but also a reduction of that explanation – the latter is required to show how the former is physically realized. More generally, just because phenomena occur in combination is no grounds for failing to distinguish them.

Greeno maintains that the morphological/systematic distinction reflects only a difference in complexity. But I think the "complicated additional analysis of how the components interact" marks an (unsharp but) important difference in kind – at least when the interactions are intricately organized. In a full blown case, it is precisely the sophisticated orchestration of the interactions that yields the explanatory insight. This element is entirely lacking (not just less complex) in morphological explanations of coffee cup or fiber optics abilities (and also in the reductive explanation of water boiling).

According to **Harré**, systematic explanations are those in which the synchronic form of an item is cited in explaining the diachronic form of its operations. This may be true, but it fails again to mention the crucial organization of the (diachronic) component interactions. Thus, the movement of a ball through a maze is diachronic, but its "form" would be explained morphologically. Harré also wants allies in his wholesale rejection of logical empiricism, particularly in his view that the deductive-nomological schema is "peculiar and degenerate." It still seems to me, though, that Newtonian explanations of planetary orbits are derivational- (hence deductive-) nomological and perfectly respectable. Admittedly, however, this schema is "most familiar" only in certain traditions, and not across the board.

Economos proposes an alternative classification of explanations: those that are heuristically convenient and intuitive, versus those that give ontologically "true coin." Her claim that the former are merely promissory notes, to be redeemed by the latter, is a metaphysical thesis about reduction. But she and I agree that my distinctions do not address that issue and can all be thought of as heuristic in her sense.

Finally, **Baron, Domotor** and **Otto** argue that the explanation of how IPSs work is derivational-nomological, but with a distinctive kind of mathematics. Computation theory, they say, plays the same role in explaining computational systems that differential equations play in physics. I see two differences. First, computation theory is not used to express a closed, complete body of empirical laws comparable to Newton's (expressed as differential equations); rather, it is a formalism for defining proper and nondefective operation of various computing systems and for determining further characteristics of systems so defined. For example, the outputs of an adding machine can be predicted (and explained, in one sense) by deriving them from the "laws" of arithmetic, plus statements of the initial conditions (inputs); but the similarity to D-N explanations is illusory. Arithmetic laws are not empirical generalizations confirmed by painstaking observation of known adding machines (or accountants); they are prior conditions that an activity must meet if it is to count as addition. If a planet fails to behave as predicted, it is not defective – the laws are (or else they are misapplied). But if a machine adds incorrectly, arithmetic is not defective – the machine is (or else it is being misused). The empirical problem is not to discover and verify computational laws, but to determine which black boxes can be interpreted as performing which computations.

Second, IPS explanation explains how something works, not why paramaters have certain values, or even why certain events occur. It is no more derivational-nomological than the explanation of how an engine works, despite the mathematical precision of its formulation. D-N explantion is subsumptive in character; we understand something because we see it as an instance or special case of what we empirically know to be invariant in nature (scientific laws). What we "see" in systematic explanation is how an organized complex of component abilities and interactions "adds up" to a new overall ability. That is not subsumptive, though, to be sure, it takes the regularity of nature for granted. Baron, Domotor and Otto have rightly pointed out that IPS explanation can employ a rigorous formal apparatus, but that does not imply that it has the more specific form of derivational subsumption.

Section 3. Several commentators noticed an oversight in my characterization of IBBs. Since the inputs and outputs make sense in virtue of a pattern they exhibit – a pattern that is extended in time – an IBB interpreter can often attribute enduring intentional "states" and changes therein. For example, a blind chess player (person or machine) must keep track of, or "remember," the current position, updating it after each move. Such attributions yield a fuller, or more complete, IBB interpretation. Some input/output patterns may be so complicated as to be unintelligible without this added completeness (thus, attributing beliefs and desires might be an essential ingredient in interpreting human behavior). But these "full" interpretations are still quite external, in the sense that they have nothing to do with internal structure or workings – they imply nothing about whether the object is an IPS, or if so, which one. They are simply further specifications of the pattern in the inputs and outputs, in the vocabulary of the interpretation.

Thus, **Dennett** points out that full IBB interpretation is all he meant by "adopting the intentional stance" (1971). **Atherton** and **Schwartz** note that you can attribute plans, hopes, and so forth, to an agent without being committed to an IPS explanation. **Arbib** implies that not all "functional components" are "structural components," and **Hayes** says that "having a particular componential structure" can be entailed by some IBB interpretations. This means, I take it, that a full interpretation might apportion enduring states among enduring "components"; unfortunately, the word "component" suggests internal structure – perhaps "faculty" would be less misleading. (As I use the term, a "functional component" must be a structural component in an actual system, though, of course, the "structure" may only be specifiable relative to another system, such as an IPS on a lower level or dimension.) **McCarthy** proposes a useful terminology for the distinction: "competence cognitivism" concerns full IBB interpretations, and a step-by-step spelling out of their cogency conditions; "performance cognitivism" involves IPS explanation of cognitive competence in terms of actual interactions of actual functional components. It is to performance cognitivism that my discussions are addressed.

Andreae balks at the claim that reasonable interpretations are "seldom hard to recognize in practice." He is thinking, however, of finding or generating interpretations in the first place, whereas the claim is only about recognizing them as reasonable, once they are proposed. **Tweney**, on the other hand, complains that reasonableness is too easy to find; we see "meaning and pattern" everywhere, even in inkblots. But none of his examples involve inputs and outputs in articulated typologies; interpreting these on the basis of observed patterns is much more difficult. **Matthews** rightly feels that the notion of "making reasonable sense" is closely related to that of "meaningfulness," but that does not argue against the one serving as a criterion for empirically testable attributions of the other. **Domotor**'s discussion of theoretical meaningfulness overlooks the difference between a meaningful theory of meaningless phenomena and an equally meaningful theory that attributes meanings to certain phenomena. It is like the difference between description and translation – only the latter involves intentional interpetation. **Arbib** is satisfied to call interpretations "judicious" and skip any "Y'know what I mean?" style definitions. Were he more candid about his own terminology, he might appreciate that all explication – scientific, philosophical, or otherwise – ultimately ends with "Y'know what I mean?" This is no excuse for shirking the hard work of clarification, nor for failing to acknowledge what remains unclarified.

Dreyfus and **Taylor** raise a more serious question. They accept my account of IBB interpretation for adding machines and mechanical chess players, but they doubt that the same goes for our "interpretations" of each other (even "applying the definitions flexibly"). Both home in on the requirement that inputs and outputs be individually and independently interpretable, once the overall interpretive scheme is specified. (This follows, as Dreyfus details, from the definition of "quasi-linguistic representation.") Taylor is right in saying that the issue is not holism (in Quine's sense); IBB interpretation is holistic, since individual interpretations are possible only relative to a successful overall scheme. The question is, rather, "contextualism": can there be an overall scheme that assigns interpretations to utterances independent of their specific contexts? Dreyfus argues that natural language may be essentially ambiguous, in that the meanings of what people actually say may be only partially and loosely constrained by the finest possible "semantics" (holistic interpretation) for speech behavior; the particular context (including the conversational situation, tone, etc.) supplies such further determinateness as there is. Taylor suggests that our sensitivity to these contexts may involve a noncognitive (in my sense), nonsemantic (in the Quine-Davidson sense) capacity that he calls "implicit grasp of a domain"; in other words, actual meaningfulness may not be fully capturable in terms of interpretations.

My views are as tentative as theirs. Cognitivism might have an escape roughly like the one I outlined for perception and action, but generalized to cover linguistic perception and speech acts. It amounts to saying that even if people are not strictly IBBs "on the surface," they nevertheless have an "inner" IBB, the inputs and outputs of which we can surmise from surface stimuli and behavior (knowing what we do about the situation, etc.). These surmised inputs and outputs would all be quasi-linguistic, and IPS explanation might proceed from there inward (**Dreyfus** appropriately mentions Fodor in this connection). At stake is whether the partial quasi-linguisticness of ordinary discourse (its "quasi-quasi-linguisticness") is to be understood as a surviving surface manifestation of a thoroughgoing, deep quasi-linguisticness. The fate of the "interpretation approach" to meaningfulness is the same issue. Or, in still other terms, it is approximately the question of how far **McCarthy**'s competence cognitivism can go, beyond what we would intuitively call "problem solving." I just do not know.

Section 4. Hayes and **Monk** share a worry about "working out a rationale"; they wonder whether this means that an IPS must "justify" its outputs to itself or "generate a method for producing" those outputs. It does not; it only means that we can give a certain kind of explanation of how the object consistently manages to produce sensible outputs, namely, it is so constructed that the internal interactions leading to its outputs will always be interpretable (by us, not it) as the steps of rationales that justify those outputs (to us). Because we see that it is so constructed, we can understand (explain) its consistent ability to make reasonable sense.

Monk adds that not every "computer-like routine" for producing correct outputs is interpretable as "reasoning the problem through." I agree; the overall IBB might have to be intentionally instantiated before its ability can be explained. For example, a computer simulation of the electronic circuit of an and-gate could itself be interpreted as an and-gate, but not (on

that dimension) as an IPS. More interestingly, holographic associative memories can be computer simulated, but the simulations are not IPSs – even though they are instantiated on IPSs. Indeed, on the infinite computer of folklore, we could presumably simulate the brain; but that would prove nothing about the explanatory form of psychology. **Charniak** says the same thing from the other direction: the failure of Cognitivism (in my sense) would not entail the impossibility of intelligent computers (though much current work would be undermined).

Cummins presses the point in a different way: there are IBBs that "cogitate" (work out rationales), but that are not explainable as systems of interacting IBB components. This does not, as he charges, render the notion of IPS ambiguous; an IPS is, by definition, a system of interpreted components. Avoiding this possible confusion is the motivation for the definition. His pills-and-liquids example is essentially similar to a diode and-gate (for which a "machine table" and "flow chart" could also be concocted). The only difference is that the temporal sequence of overall physical states in his device more closely parallels the order of boxes in the chart. But he is quite right that since they are overall states of the whole object, rather than interacting states of separate IBB components, the object is not an IPS (in my sense); it is a physically instantiated IBB. In effect, the machine table that he provides is a competence theory (in **McCarthy**'s sense), and is the sort of "fuller" IBB interpretation that I allowed earlier.

Section 5. Arbib and **Hayes** are bothered by intentional instantiation, and the dimension/level distinction. Arbib supposes that such precise terminology rules out typical computer programs as IPSs, which is a simple misunderstanding; instantiation in a program written in another "language" is not incompatible with being an IPS, though, as just mentioned, not every computer instantiated IBB is (on the same dimension) an IPS. Hayes wonders how operations on numbers and labeled trees can be included in the interpretation of a chess-playing IPS, since they do not pertain to chess. But on that dimension the representations are of position values, possible move sequences, and the like; it is only when interpreted as such that their relations and transformations make sense in the context of explaining how a chess player works. The interpretations of them purely as numerical or abstract tree-search operations are on other (instantiating) dimensions. It is because intentional instantiation is (by design) so often transparent, that its conceptual importance has been overlooked.

Economos and **Hayes** also miss the point about physical instantiation, she saying that significances themselves must interact (or "have causal consequences"), and he that these interactions are symbolic and informational, rather than causal and energetic. On any intentional dimension of description, of course, this is quite correct (whether or not to use the term "causal" on such dimensions is a verbal issue). But the power of the physical instantiation idea is that the very same interactions can simultaneously have physical descriptions – indeed, they must in any concrete, working system that we know how to construct (compare **Maxwell**). The physical system is designed in such a way that in obeying physical laws it will, when interpreted, also "automatically" obey what Hayes calls "the very rules which define the meanings of the symbols" (this being just my condition on cogent interpretability; see also Davidson, 1970). If Economos thinks this still leaves a mystery, or violates the laws of thermodynamics, then she must think the same about pocket calculators.

Greeno and **Simon** believe the possibility of reduction shows systematic or IPS explanations to be weaker than those of physics. There is no dispute if "weaker" means "less-fundamental," in the sense that "fundamental" physics is (by definition) that to which all other explanation is supposed to be reducible. Nor is there dispute if they mean that an explanation and its reduction together explain more than the explanation without the reduction. But if, as I suspect, they mean that IPS explanations are only a temporary stopgap, until more powerful physical explanations can come in and take over, then I disagree. The laws of physics do not and cannot say anything about chess; they can no more explain how a chess player plays chess than the additivity of lengths can explain how a slide rule multiplies. In general, reducing explanations explain different things than the explanations they reduce.

Two final points: Simon (like many people) regards reducibility as essential to Cognitivism and he concludes that psychophysical bridge laws are necessary after all. The conclusion does not follow. "Bridge law" is a technical term in a very specific account of how some reductions proceed, for example, the reductions of classical thermodynamics and optics. I have outlined a general account of reduction that includes the bridge-law procedure as a special case and systematic reduction as a quite different special case; they should not be confused. Greeno seems to think the following five contrasting pairs denominate approximately the same distinction (among explanations): systematic versus derivational-nomological; qualitative versus quantitative; informal versus formal; semantic versus syntactic; and less testable versus more testable. I would not comfortably assimilate any two of these, but let me just say that IPS explanations can be expressed quite formally (computation theory) and be fully testable (do the interactions occur as required or not?) and yet be systematic and, in some sense, "semantic."

Section 6. Baron, Dennett, and **Hayes** kindly remind me of the "rightly" in my remark about hologram models: "neurophysiologists have . . . rightly confined their speculation to recognition and recall processes." Baron adds a stern and generally stifling warning: "One should not think of the holographic models as any more than they were intended to be. . . ." His own further remarks remain limited by this injunction; thus, he discusses the possible difference between static and dynamic models only with regard to memory.

Dennett puts the genuinely hard question more directly: can we even imagine something roughly hologram-like "getting information processed in a much more dramatic sense" than mere "content-driven" association? My vague outline of a chess-move generator was meant to suggest that we can (or, at least, that it is not obvious that we cannot); unfortunately, however, it is at best "suggestive." As Dennett must realize, it is difficult to avoid prejudicing the issue, just in formulating it. Are we asking for something that "processes information" but is not an "information processor"? That sounds incoherent. Yet we want something that does more than simply transmit the "information" (to another place or time) or "translate" it (into another representational mode). Even sorting it, selectively distributing it, or having it control a mechanical process would not be sufficiently impressive. We want a device that "uses" the various things it knows to make intelligent overall choices, draw intelligent net conclusions, and so on. We now know that IPSs can do this, at least to some extent, but is it only IPSs that can do it? The trouble is that if we redescribe "behaving intelligently" as "getting information processed in a dramatic sense," then our terminology tends covertly to beg a deep question about the nature of intelligence. The central aim of my discussion is to define "information processing system" with enough precision to make clear that being intelligent does not entail being an IPS. The guiding thesis of Cognitivism is not a conceptual truism, but an important empirical hypothesis. Dennett seems to appreciate this. If, after that, he still finds it difficult to envision specific, detailed alternatives ("What else could it be"?), it is not surprising – our intellectual and technological heritage leaves us all in the same boat.

The foregoing should also explain my lack of sympathy when **Greeno** says he would "prefer to have a broader concept" of IPS or when he objects that holograms "clearly" process information or when **Hayes** claims they instantiate "a kind of computation."

We need not haggle over particular words, but we must recognize that narrow, refined concepts are at least as important to scientific progress as broad ones. GPS, SHRDLU, DENDRAL, CHESS 4.5, and the like all exhibit a certain kind of structure that holographic associators, string net analogs, and maybe people do not exhibit. I hope my definitions capture it, and I want a term for it. Without such a "narrow" term we cannot even sensibly ask whether human intelligence is realized in that kind of structure, or whether there are limits to what can be so realized – and those are crucial questions we all care about.

The trouble with the language of "information theory," "control theory," "cognitive theory," "computation theory," "decision theory," and so on is that it is used in a variety of technical and intuitive ways. The intuitive senses make it seem obvious that memory is "information storage and retrieval," that perceptual recognition is a "transformational computation," that making up your mind is a "decision process," and that reveries, hunches, inclinations, and longings are "cognitive states." Then one or another technical sense makes it seem like something nontrivial and scientifically important has been said. To take just one example, **Baron** speaks of "the very complex way that information is routed and processed within the brain"; what is he talking about? If he means "information" in Shannon's sense, then, of course, we can regard the brain as routing and processing it, just as we can the blood, ecosystems, and telephone networks. If he means "information" in the sense that Sherlock Holmes could get a lot of it from amazingly subtle clues, then he knows little more than did William James, or even LaMettrie, about what that has to do with the brain. But if he means "information" in the sense in which chess programs route and process information about pieces, options, position values, and so on, then transferring the notion to people is conjectural, and, in the present context, question begging. More likely, however, he has jumbled all these senses together, and thus cannot imagine doubting that his phrase describes the decision processes in Sherlock Holmes's neurons. It is time to start using our terminology more selectively so that we can say exactly what we mean and see what our evidence really supports.

Section 7. Monk feels that the seriousness of the hurdles is hard to assess because the proper domain of Cognitivism (or cognitive psychology) is not sharply delineated. Fair enough. The extent of the domain is itself an empirical question to be settled, in effect, by how much turns out to be cognitively explainable. The threat of the hurdles, if any, is that the domain will prove disappointingly small. Hence, my strategy is to suggest that each hurdle might be noncognitive and yet infect large areas that one would hope to find in the cognitive domain. Thus, **Domotor's** dismissal of them as "kinky exceptions" (even physics has exceptions, he says) entirely misses the point. **Tweney** supposes that the problem in all three hurdles is the role of consciousness. That strikes me as most plausible for understanding, least for skills. But, frankly, I did not mention it in any of the cases, because I do not have the foggiest idea of what to say. I have so little grip on the notion of "consciousness" (though I do not deny that it is genuine and important) that I cannot imagine what a theory would have to do to account for it, nor do I know what a theory would lack if it did not. Call it a fourth hurdle, if you like, but I cannot even see what "clearing" it would involve.

On moods. Charniak objects that moods are not so pervasive as I make out: they have minimal effect, he says, on currently important problems, like how we do pronoun reference, or tell telephones from briefcases. I am not convinced. Consider a moment at a party, when someone says: "This is just too much" or "He's driving me up the wall." Might not the objective referents (and/or the linguistic antecedents) of "this" and "he" depend importantly on whether the speaker is irritable, high, hilarious, or feeling frisky? Understanding the sentence requires understanding the situation, which includes understanding the people in it – their moods, their personalities, their reactions to themselves and to one another. And that understanding itself will be a function of the mood of the hearer.

We should step back, however, and ask why pronoun reference and telephone/briefcase discrimination are "currently important problems." There are two sandwiching reasons: first, they are very difficult (as yet unsolved); second, according to current views, they ought to be comparatively straightforward and tractable – compared, that is, to full discourse analysis and appreciation (or adequate "representation") of genuine human situations. These latter will almost surely require attention to moods. (What is the import of this adjective or simile? Is that sympathy, triumph, or bemusement in her smile?) Relative mood-independence may in fact be part of the reason that certain grammatical and recognitional problems are currently important; but if so, that is just a backhanded acknowledgment that current theory is stumped by moods.

Pinker is seduced by "the full power of a Turing machine" and thinks the problem with moods must lie in finding not just any old account, but a "principled or elegant" one. Now it may be that an infinite Turing machine could calculate every movement of every molecule in a person's body, given the initial conditions and physical stimuli. But that has nothing to do with psychology – it could do the same for Lake Michigan. To say instead that it could calculate certain behavioral outputs in terms of "estimates of subjective probabilities, assignments of salience to [input data]" (*not* stimuli!), and so on, is to assume the adequacy of competence cognitivism for that range of behavior. But it is not obvious that vindicating this assumption for mood-influenced behavior would be "routine if tedious" (see **Dreyfus, McCarthy,** and **Taylor**). Indeed, **Pinker's** own "third and more ominous possibility" implies as much: behavior affected by moods might not be fully determined by cognitive states alone. My concern, however, is with performance cognitivism, and in that case the full power of arbitrary Turing machines is quite beside the point. The question is the "actual mechanism" by which the behavior is produced, and whether that mechanism can be consistently interpreted as an IPS.

There is a problem of cross purposes when **Atherton** and **Rey** object that we do "justify" moods. Certainly, we explain the occurrence of moods, and, in that sense, give "reasons" for them; thus: "Naturally Jane is upset; look at all that's happened to her." Also, we criticize them as unseemly or overwrought ("Not that much happened to her.") Similarly, we sometimes explain beliefs by citing reasons for their occurrence, such as: "He thinks it's a trap because he's paranoid," or "She only believes it because she wants to so much." But there is a different sense in which we more often "explain" beliefs by "giving reasons": we cite arguments from available evidence to the effect that the belief is (probably) true. Likewise, we support desires by offering arguments to the effect that what is desired is desirable, given evidence and other desiderata. I do not believe we give reasons in this latter sense for moods, and that is what I mean by saying we do not "justify" them (roughly, "epistemically"). Moods do not follow from evidence or premises and thus they are never correct or incorrect as inferences; at most they are understandable and proportionate, or peculiar and disproportionate. Moods do not "represent" the world, either rightly or wrongly, as being one way rather than another; at most, they affect the way the world is represented, and how one behaves in it. The most common analogy is rose-colored glasses, which do not "depict" the world as rosy (either justly or unjustly); at most, they infect and influence one's judgments about the "colors" of things.

By the way, some emotions might be different from moods in this regard: envy, resentment, and gratitude, for example, seem more likely to have a "representative" or "belief-like" component than do good cheer, boredom, or grouchiness. That is, there seems to be an aspect, or even constituent, of them that

is either true or false, and likewise either rationally justified or not.

Baron, McCarthy, Natsoulas (maybe), **Pinker, Rey,** and **Rorty** all seem to assume that whatever is noncognitive about moods is ipso facto physiological – thus their allusions to drugs, brain chemistry, and voltage changes. It is as though Cognitivist and physiological explanations were the only conceivable candidates (with the former exhausting psychology). I confess it astounds me how quietly and easily this gets taken for granted, without a hint of an argument, and usually without even being noticed. Moreover, it strikes me as an intellectual cop-out: if *our* psychology cannot handle it, then no psychology can, and we will just imagine some magical "chemistry" that does – for example, the "dormitive power" of opium. Only McCarthy is candid about it: melancholy is "just a high concentration of *melancholine* in the blood" (and bilious tempers in general have to do with bile of various colors? ... or is it psychedelic voltages?) And tacking on a high-sounding remark about altered sentence probabilities, changed coupling parameters, different functional modes, reallocated processing capacity, or program/protoplasm interaction is just shriller whistling in the dark. Nothing specific about moods is accommodated, except that they vary, and make a difference.

Simon, however, has an interesting idea: a mood may not be an "integral quality," but rather a "complex interplay" of numerous cognitive states (together, perhaps, with "feelings"). It would be like an exciting or depressing chapter in a novel; the characterization does not apply to any particular sentence or to any specifiable structure of sentences, and yet there is nothing on the pages but ordered sentences. Thus, even if boredom is no particular cognitive state, it might be an overall feature of a large group of them. Still, suggestive as the proposal is, several questions come to mind. First, moods *are* in a sense "integral" or "coherent"; we do not find ourselves one-third bored, one-third fascinated, and one-third crestfallen (on the other hand, it is common to be both excited and nervous, or both serious and aggressive). In a story, the mood coherence is dependent upon the author's prior sense of real life and is appreciated by an already sensitive reader, but a psychological account must explain the "coherence" of the "original." Second, one's successively occurring new thoughts, inclinations, and so on (whether precipitated by perception, conversation, rumination, or whatever) most often "fit" the mood one already has. How, if a mood is just an overall character of many separate cognitive states, can it have this general net influence on subsequent states? Third, moods affect how things "seem"; a mosquito can be anything from a nuisance to intolerable; the same flowers or music can range from deliciously sweet to negligible to cloying. And fourth, in different moods, the flesh hangs differently from the bones: facial expressions, "body language," tone of voice. How is the model of moods in a novel going to account for these?

I do not say that these questions cannot be answered, or even that "chemistry" will not be just what Cognitivism needs to complete the answers. I think many commentators feel that my questions do not prove anything – that I have failed to give convincing counterexamples and am just waving my hands. So they wave their hands back, and think we are even. But that misses the point. Cognitivists have powerful and well-developed stories to tell about problem solving, sentence parsing, memory organization, and the like. Whether ultimately right or wrong, these are honest, articulate efforts, sensitive to the details of the phenomena, with considerable explanatory content and initial plausibility. But when it comes to moods (and, to a lesser extent, skills and understanding), practically all they have is hand-waves: dormitive power, with a little decorative jargon. This situation may or may not last, but the longer it does, the more likely it is to be important. If and when the frustrations and limitations of AI and cognitive theory begin to seem like anomalies and counterevidence (a different point for each person), then "hurdles" such as these are good places to look for the underlying problem. Hand-waving back at me does not obviate my questions; it underscores the need for them.

On skills. Baron and **von Glassersfeld** effectively take the Cognitivist line to which I allude: in their words, we automate, or store for later invocation, (pre-) motor patterns, which we have previously built up or consciously generated. They say nothing, either directly or indirectly, about why paying attention interferes with skills, or why articulate people are so often inarticulate about them (**Rey** "sees no reason" why something could not be said, but he does not say it). More important, they do not in the least confront **Dreyfus**'s difference-in-kind objection. How can they purport to answer my worries, when they do not even address them?

Arbib and **Hayes** believe human skills have already been computer-modeled, with Samuel's checker player and Stanford's DENDRAL. They might as well have stuck with the chess example, but then the response would have been more obvious: CHESS 4.5 explicitly evaluates hundreds of thousands of options per move, which (on the best evidence) is three orders of magnitude more than any person. A program that matches human competence does not necessarily model human performance [cf. Pylyshyn et al., *BBS* 1:1. 1978]. Of course, you can hypothesize unconscious (and, so far, undetectable) human performance, comparable to what the program does, but then again you face the Dreyfus objection. Furthermore, known IPSs have actually matched only a narrow range of human competence – always for specific, highly structured ("toy") problems (spectrum analysis, disease diagnosis from given symptoms, mathematical routines, rule-defined games). Just suppose that human performance is in fact realized in a radically different way. Then the possibility even of competence matching might be fundamentally limited. This is plausible enough, given what we know: how many computer programs can do "childishly easy" things, like watching a movie and telling you who is angry, who the good guy is, and who won?

Rey rightly points out that the distinction between conscious deliberation and whatever is unconscious need not coincide with that between cognitive and noncognitive (even if deliberation is a pretheoretical paradigm of cognition). I also agree that the possibility of consciousness is more "remarkable" and in need of explanation than that of nonconscious information processing. My prima facie point is only that, on the level of reflective common sense, our archetypically cognitive and skillful capacities are very different. So a theory that proposes to assimilate the latter to the former on a deeper level is driven to postulate something that runs counter to the evidence of common sense – which is perfectly respectable in science when the postulate is satisfying on other grounds. But in this case, the postulate has little to recommend it except that the theory would need it to remain viable. It fits further available evidence, at best spottily and poorly; and when you press for details, it assumes the proportions of a *deus ex machina,* so powerful that you wonder why it does not take over completely. If unconscious chess "processing" is really just as cognitive (working through rationales) as conscious counting out, the only operative difference being a thousandfold increase in efficiency, then why does that last tenth of a percent have to be worked through so clumsily and slowly? And why are so many less rule-structured, but otherwise "easier," skills so much harder to program at all?

Simon cites H. Simon's work on different sized "chunks," for short-term memory and processing (e.g., in chess play), but completely ignores the skilled "preprocessing" by which the relevant chunks are selected. **Puccetti** suggests this preprocessing might be conscious, even though we are unaware of it, and he calls it "cognitive," even it it does not fit the IPS or "computer" model. I think we mainly agree, though I find his terminology foreign. The split brain work is intriguing, but I have no view on it. I am sympathetic when **Rorty** and **Schwartz** note a deep similarity between understanding and at least some skills, but I do

not see the sharp distinction that Rorty wants between these and other skills. Typing from a manuscript, driving in traffic, and the like, even when quite "automatic," seem to involve some degree of understanding, and I suspect that a lot of light conversation does not involve much more. I think Rorty's distinction is another holdover from the Rylean "intellectualist legend" that Schwartz mentions.

Finally, **Hayes** and **Taylor** point out an important feature of many skills: acceptable performance is constrained by justifiability. Thus, no matter how people actually generate chess moves, the moves are good precisely to the extent that they can be rationally justified – and good players would endorse the rationales (after studying them). Hayes thinks this "argues strongly" that the experts were working through such rationales all along, but I see no argument stronger than: "What else could it be"? (Cf. **Atherton.**) Taylor has a name for what else it could be: "implicit grasp of a domain" with its "own kind of rationality." I believe that expressions like this are a valuable contribution to discussion, and reflect a sensitive feel for the phenomena, but (like everybody else) I wish we had a more worked-out account of the nature and physical instantiability of these phenomena.

On understanding. McCarthy conjectures that the difference between merely manipulating symbols according to rules and understanding them is whether they have been translated into an "internal language." In the Cognitivist view, however, such a language is just so many more symbols manipulated according to (internal) rules. **Schwartz** sees the difficulty in internal intentional states "making sense to the organism or mechanism using them, not just to us." And **Matthews** is worried that Cognitivism will "impale itself" on this problem; translation (the interpretation approach) cannot ultimately handle it, because of a regress threat.

But it seems to me that one strength of Cognitivism is its prima facie ability to answer such questions: understanding "mentalese" just corresponds to manipulating the internal symbols according to appropriate rules. Precisely what the IPS model shows is that this is possible without an extra "little man in the head" to interpret the symbols and apply the rules; gone is the bugaboo of the question-begging homunculus. In this sense, a chess playing machine understands perfectly well the moves we type in, and also many of their implications.

Something else is bothering me: why do some rules make sense, and others not? For surely there are sets of rules that we or a machine could master, and yet that make no sense at all; a versatile ability to "apply" them would not be tantamount to understanding the "symbols" manipulated, because there is nothing to understand. It is tempting to suggest that the rules we actually follow make sense "to us," just because they are the ones we follow. But how, then, can our rules change, where a nontrivial necessary condition on the new ones is that they make sense? It might be answered that our "real" rules do not change; all that change are the more specific forms that we "translate" into them. Or perhaps (as **Rey** may have in mind) there are unchanging meta-rules that determine sensible rules.

But these unchanging (eternal?) rules must be not only "built-in" but also astonishingly powerful. The rules of chess would not suffice because they are not general enough. Yet the abstract structures of set theory, or even Turing Machines (which are very general, when you allow intentional instantiation), are not concrete enough: all sorts of nonsense systems can be instantiated ("modelled") in set theory. The hypothesized built-in rules must be both so general that they cover all possible genuine cogency conditions and, at the same time, sufficiently detailed and specific to rule out millions of possible sets of nonsense rules. That is, directly or indirectly, they must fully specify all and only the appropriate manipulations of any symbols that pertain to any (humanly) intelligible subject matter whatsoever (and no others). Since appropriate manipulations often depend on individual symbols (what they "mean"), I suspect that this double requirement forces one back to a Fodor-like universal and innate "language of thought" (or Leibnizian "universal characteristic"), in which anything that any human being could ever say or imagine is already unambiguously expressible. But one person's *modus ponens* is another's *modus tollens*. I find the conclusion so incredible that if Cognitivism implies it, then so much the worse for such premises.

It seems to me, however, that there is a deeper side to this: understanding pertains not primarily to symbols or rules for manipulating them, but to the world and to living in it. Linguistic articulation can be a vehicle for such understanding, and perhaps articulateness is prerequisite to any elaborate understanding. But cases where facility with the symbols is plausibly sufficient – like well-defined games, mathematics, and AI "micro-worlds" – are very peculiar and, I think, parasitic. Paradigms of understanding are rather our everyday insights into friends and loved ones, our sensitive appreciation of stories and dramas, our intelligent handling of paraphernalia and commerce. It is far from clear that these are governed by fully explicable rules at all. Our talk of them is sensible because we understand what we are talking about, and not just because the talk itself exhibits some formal regularities (though that, too, is doubtless essential).

When the rationalists took cognition as the essence of being human *(res cogitans)*, they meant especially theoretical cognition, as in mathematics and mathematical physics. The understanding manifested in arts and crafts was not a different phenomenon, but just imperfect theory, sullied by obscurity and confusion. Cognitivism is heir to this tradition: to be intelligent is to be able to manipulate (according to rational rules) "clear and distinct" quasi-linguistic representations – sullied now by omissions, probabilities, and heuristics. But, deported from the immortal soul, they forfeit their original anchorage in God's honesty and the natural light of reason. Bereft of their credentials from above, the distinction of certain procedures as "reasonable" floats adrift, unless it can be otherwise explained. Evolution comes vaguely to mind, but more needs to be said. My own hunch is that the intelligibility of rational "theorizing" (from unconscious "cognition" to set theory) is a derivative special case of an antecedent, atheoretical, intelligent "practice" – a prior "grasp" of how to get along in a multifarious existence. If articulate theory is one developed derivative, there can be others: the appreciation of fine art, a subtle sense of personality, the "mastery of metaphor" (**Rorty**, from Aristotle), even creativity and wisdom. We will understand understanding when we understand its many forms, primordial and refined. In the commerce of understanding, words are only money.

Section 8. Atherton and **Dreyfus** (and, in a different way, **Harré**) point out that part of the attraction of Cognitivism lies in preconceptions about what counts as "explanatory" or "scientific" (see also Dreyfus, 1972, pp. 143–46). I think this is deep and important. While Cognitivism has broken away from the derivational-nomological mold of physics and a few other disciplines, it retains a larger scientific prejudice favoring complexity over richness and elaboration. The ideal is a set of clearly delineated parts or factors, which combine in exactly specifiable ways, such that the complexes are precisely the phenomena to be explained. It is a kind of atomism, whether of components or variables, and its pedigree is coeval with that of clear and distinct ideas. The sort of understanding that Cognitivism is least able to account for may be just what is needed in psychology itself. **Harré** cites the "ethogenic movement," which draws its "rhetoric" from literature and drama (rather than mathematics and technology?). Whether this would mean psychology could not be "scientific" strikes me as largely a verbal question; it could certainly be empirical, convincing, and illuminating. But, as I said to **Taylor** (and admitted earlier to **Dennett**), what I mostly feel the lack of is a detailed account of how such things could be instantiated. Here Cognitivism has (so far) all the ad-

vantages; my hologram speculation was at best a vague indication of what an alternative might possibly look like.

Still, I am not at all happy with **Charniak's** and **Hayes's** invocation of Kuhn to blunt criticism. In caricature, the sentiment is: don't bother us with problems – "normal scientists" are supposed to be pig-headed. That is not so, and Kuhn does not say it; and, of course, nor do Charniak and Hayes quite say it. Charniak says a critic must "become a specialist" who can "sit down with researchers and go over computer listings," and perhaps also come up with a "plausible replacement." Hayes suggests that Cognitivism has not yet "crystallized," and hence expressing misgivings is "antiscience." The implication in both cases is that I am out of place to question and should rather keep my mouth shut.

Not surprisingly, I disagree. Paradigms are no more sacrosanct than anything else in science. I am not sure what it means to say that Cognitivism has not crystallized; but if it has not, maybe that is because it is not going to – it has probably already had more genius-hours invested in it than has the whole of classical physics. I agree with Charniak that in this kind of debate there are few clear-cut rebuttals; short of a stunning new achievement, the best a dissenter can do is "wind a web of dissatisfaction." For that the prerequisite is a well-informed and penetrating overview, not the nitty-gritty of workaday expertise. The aim is not to pinpoint the flaws in particular efforts, but to begin articulating what is common in a variety of well-known nagging frustrations – to waken, it is hoped, a few from their dogmatic slumbers and perhaps, somewhere, to stimulate "a new Copernicus."

ACKNOWLEDGMENT

I am indebted to Jay Garfield for valuable assistance in formulating these replies.

Call for papers (and topics, authors, nominations, and suggestions)

The Behavioral and Brain Sciences is now calling for papers (as well as for recommendations of topics and already published articles or books) for *Open Peer Commentary*. Respondents are requested to provide an explicit rationale for seeking or recommending *Commentary* and to include a list of possible commentators (see Instructions to Authors and Commentators).

We are also accepting informal nominations for (A) Corresponding Associate Commentators (who perform *Open Peer Commentary* on accepted articles) and for (B) members of the Board of Editorial Commentators (who referee submitted manuscripts in addition to performing *Open Peer Commentary*). A list of those who have thus far joined the Associateship hierarchy is available from the editorial office (and will appear in Volume I, Number 4 of this Journal). Qualified professionals in the behavioral and brain sciences who have either (1) been formally nominated by a current BBS Associate, (2) refereed for BBS, or (3) had a commentary or article accepted for publication can serve in capacity A or B.

To help in optimizing the service of *Open Peer Commentary* to the behavioral and brain science community, the editorial office would welcome suggestions from readers and Associates as to the optimal (1) length of a commentary, (2) number of commentaries per treatment, and (3) format for treatments.

Communications regarding these matters should be addressed to the Editorial Office, Behavioral and Brain Sciences, P.O. Box 777, Princeton, N.J. 08540.