Having Thought

Essays in the metaphysics of mind

John Haugeland

HARVARD UNIVERSITY PRESS
Cambridge, Massachusetts and London, England 1998

Contents

Toward a New Existentialism

Understanding—making sense of things—is the mark of the mental. This is not to deny that intentionality, rationality, objective knowledge, or self-consciousness might also be marks of the mental, but only to put the emphasis more nearly where it belongs. For, in my view, each of these latter, properly understood, presupposes understanding and is impossible without it. Understanding is always "of" something—objects, in a broad sense—but this of-ness is not the same as that of beliefs and desires. Thus, understanding is not the same as knowledge, a special kind of knowledge, or even a complex structure or totality of knowledge. Rather, understanding is a fundamentally distinct phenomenon, without which there could be no knowledge or mind at all. It needs, therefore, a different discussion—a discussion which, as it seems to me, has been missing in philosophy.

Understanding is the mark of the human. This is a better way to make the point, and for two reasons. On the one hand, understanding is *not* exclusively *mental* but is essentially corporeal and worldly as well; but, on the other, it *is* exclusively (and universally) *human*. Accordingly, intentionality, rationality, objective knowledge, and self-consciousness, properly understood, are likewise exclusively human. By 'human', I don't mean specific to homo sapiens. Humanity is not a zoological classification, but a more recent social and historical phenomenon—one which happens, however, so far as we know, to be limited to homo sapiens.

It is, of course, tendentious to deny intentionality and rationality to other mammals (never mind to prehistoric homo sapiens). That there is a difference in kind, not merely in degree, between the "cognitive" capacities of people and those of other mammals strikes me as so obvious as to need no argument; and I will offer none. No doubt,

there are senses in which many animals can be said to "know", "want", and even "understand" things around them; but these are not the same as the senses in which people can be said to know, want, and understand things. Classing these uncritically together is as great an obstacle to insight as was classing whales with fish or the sun with the planets. Disputing the terms in which to express this is pointless.

More interesting is the question of *what* distinguishes people from non-people: what—if anything—is the root or essence of their distinctiveness. Many answers have been offered, from being made in God's image, or having rational or immortal souls, to the capacity for language, culture, and/or free recognition of normative constraints. In my view, the last of these comes closest—indeed, is exactly right, given a certain reading of "free recognition". Again, however, I do not undertake to defend this alternative against others, but at most to begin to articulate it.

Existential commitment is the mark of the human. This is an even better way to make the basic point, for, as it seems to me, human understanding is itself made possible by the distinctive sort of commitment that I call *existential*. It is the capacity for this sort of commitment that I am inclined to think is relatively recent—almost certainly more recent than language, and perhaps more recent than cities and writing. Like city-building and writing, the possibility of existential commitment is part of a cultural heritage (not just a biological or "natural" capacity). But, though and as culturally born and harbored, it is precisely a capacity for *individual* freedom: the freedom, namely, to take responsibility for the norms and skills in terms of which one copes with things. The ability to take such responsibility, to *commit*, is, as I attempt to show, the condition of the possibility of understanding, hence of knowing, objects.

These ideas are not new. They are announced, if not emphasized, in Kuhn (1962/70), and developed further, though rather differently, by Heidegger (1927/62). What I mean by 'existential commitment' is closely related, so I believe, to what Heidegger meant by 'authentic care', and also (albeit less closely) to what Kierkegaard meant by 'faith' and Nietzsche by 'autonomy'. A philosophy of mind and of science in which these essentially human capacities are restored to center stage is what I mean by "a new existentialism". But the point is not limited to intellectual pursuits. The general form of free human commitment—or care or faith—is love. Thus, best of all:

Love is the mark of the human.

THE THIRTEEN ESSAYS collected here, spanning some two decades, are all about understanding and intelligibility in one way or another, often several. They are arranged, roughly by topic, under four heads: Mind, Matter, Meaning, and Truth. As it turns out, this arrangement is also roughly chronological.

Under the first head, MIND, stand three essays from the late seventies, two about cognitive science (or artificial intelligence) and one about Hume. "The Nature and Plausibility of Cognitivism" (1978) is concerned with understanding in two complementary ways. How, on the one hand, can the mind itself be understood scientifically—in particular, what is the structure of the scientific understanding sought in cognitive science? And how, on the other hand, could a mind, so understood, itself be an understander? The main conclusions (which I still take to be basically correct, even if, in retrospect, awkwardly developed) are: first, that, though the character of the explanatory grasp sought in cognitive science is scientifically unprecedented, it is nevertheless perfectly legitimate; but, second, that the systems so intelligible are themselves incapable of understanding anything.

"Understanding Natural Language" (1979) pursues the latter theme, arguing that, even in the special case of understanding prose—a case particularly congenial to AI—no system lacking a sense of *itself* as "somebody" with a complete life of its own (and about which it particularly cares) can possibly be adequate as a model of human understanding. I call this cared-about wholeness "*existential* holism", and offer a number of examples to illustrate its importance to ordinary language ability.

"Hume on Personal Identity" (~1977), the oldest and shortest essay in the volume, is also about the wholeness (or unity) of the self, but in the limited context of an exegesis of Hume's recantation of his own earlier account in the appendix to the *Treatise*.

The essays under the second head, MATTER, address putative constraints on the *intelligibility* of mind in nature—particularly in its relation to the material or physical. "Analog and Analog" (1981) considers and rejects the too-common thesis that any analog system (for instance, a brain) can be digitally simulated to any desired degree of precision. The principal contribution is an analysis of the notions of digital and (especially) analog devices, in terms of which the thesis can so much as be responsibly confronted.

"Weak Supervenience" (1982) challenges that version of supervenience-based materialism that is equivalent to token identity theory,

and proposes a substitute "weaker" version (now usually called "global supervenience"). The paper first rebuts Davidson's alleged proof of the token identity of mental with physical events; then shows that weak supervenience does not entail token identities; and, finally, presents some examples meant to suggest that token identity theory is in fact rather implausible.

"Ontological Supervenience" (1984) extends that implausibility argument (in a somewhat irreverent tone) by articulating and undermining a handful of seldom-explicitly-stated "intuitive" considerations that might seem to support a materialist identity theory.

Under the third head, MEANING, the chapters are at first glance more diverse; but they are all concerned with that relationship, whatever it may be, between us and the world, in terms of which we can be said to have minds and be intelligent at all. "The Intentionality All-Stars" (1990) uses the various positions on a baseball team as a whimsical metaphor to sort and relate the most common contemporary approaches to the old problem of intentionality. Three positions are examined and contrasted in particular detail: (i) the idea that intentionality resides primarily in language-like internal representations, in virtue of the processes that use and modify them; (ii) the view that intentionality resides primarily in situated agents, in virtue of the patterns of interactions between such agents and their environments; and (iii) the suggestion that intentionality resides primarily in the social practices of a community, in virtue of the instituted norms sustaining and governing those practices.

At the time of that writing (as late as 1988), I attributed this third suggestion to Heidegger, Dewey, Sellars, and Brandom (among others); and I cautiously endorsed it myself. It now seems to me that the attribution to Heidegger (at least) was quite mistaken; and, what's more, my own view is now significantly changed (though I still think mine and Heidegger's are a lot alike).* There's no denying that social institution and its norms have been critical to the emergence and maintenance of human culture; and, so, to that extent, they have also been prerequisite to what I now regard as essential to genuine intentionality: human understanding and commitment. (There is, to be

* In my "Heidegger on Being a Person" (1982), on which the relevant section of the All-Stars was originally based, I attributed to Heidegger and tacitly endorsed the thesis or "slogan": *All constitution is institution*. (18) I now repudiate both the attribution and the endorsement; Brandom, however, still embraces the idea.

sure, an intentionality-*like* phenomenon for which social norms alone are sufficient, much as there is one for which biological-functional norms suffice.) But existential commitment is crucially *not* social; and, as such, it makes possible a kind of normativity that goes beyond anything merely instituted. (See especially chapter 13 below.)

"Representational Genera" (1989), perhaps the most disheveled piece in the volume, undertakes to distinguish qualitatively different *kinds* of representation—not just species, but genera—on the basis of the characteristic structure of what they represent. The motive for the project, at best partially realized, is to ask and determine what might be distinctive of so-called distributed representations (the kind that, apparently, would be easiest to implement in networks of neurons). Along the way, however, a lot of effort is expended in setting up an approach to the problem, and illustrating it in terms of the more familiar cases of symbolic and pictorial representations.

"Mind Embodied and Embedded" (1995) argues, from principles of intelligibility drawn from systems theory, that the customary divisions between mind and body and between mind and world may be misplaced, in a way that more hinders insight than promotes it. The suggestion is that trying to understand the structure and functions underlying intelligence in terms of interactions across mind/world and mind/body "interfaces" might be like trying to understand the operation of an electronic circuit in terms of divisions that arbitrarily cut across its *electronic* components. That is, mind, body, and world might not be the right "components" in terms of which to understand the operations of intelligence. Meaning may be as much a corporeal and worldly phenomenon as it is "mental".

The four essays under the fourth head, TRUTH, have more in common than do those under any of the earlier heads. All four are concerned with the possibility of *objectivity*, and they all approach it in terms of an idea of *constitution* grounded in commitment. "Objective Perception" (1996—though written several years earlier) argues that, in order to specify the object of *human* perception—a kind of objectivity not available to animals—the object itself must be constituted in terms of constitutive standards to which the perceiver is antecedently committed. It is also argued that such commitment does not (at least not in principle) require language.

"Pattern and Being" (1993) brings that same point about constitution to bear on Dennett's "mild realism", as propounded in his "Real Patterns" (1991), arguing that his central discussion of patterns is con-

fused unless a systematic distinction is drawn between two different levels of patterns, both of which are required for the reality (or being) of entities, in the sense he intends. Once that distinction is in place, moreover, his trademark notion of a "stance" can be made considerably clearer, and the intentional stance, in particular, can be divided into two distinct versions: a weaker one for animals and computers, and a stronger one for people—only the latter of which involves understanding, and thus has properly to do with intentionality.

"Understanding: Dennett and Searle" (1994) undertakes the unlikely task of reconciling Dennett and Searle on the prerequisites for genuine intentionality, by agreeing with each on a number of his most cherished views, while disagreeing with each (sometimes both) about a few points that strike me crucial. The pivotal issue is *understanding* (as distinct from mere knowing or believing), which, as I read them, neither Dennett nor Searle seriously addresses, and, without which, neither of their accounts of intentionality can be adequate. In the course of the discussion, I defend a sort of compromise on the disputed cases of animals and AI systems, by assigning them together to a new category—systems with *ersatz* intentionality—thereby preserving the best intuitions on both sides.

"Truth and Rule-Following" (new in this volume) is the longest and most difficult chapter. The aim is to spell out more thoroughly the fundamental ideas of constitution, commitment, and objective understanding introduced in the preceding three chapters, and to show how they enable a new account of truth in terms of beholdenness to objects. The principal innovations are an explicit distinction between norms of proper performance (such as might be socially instituted) and those of objective correctness, and the concept of an excluded zone—which shows for the first time how empirical beholdenness is concretely possible. Interesting corollaries include: (i) a distinctive exposition of the interdependence of objectivity with subjectivity, via the free commitment to standards that grounds objective constraints; (ii) an alternative to coherence theories of truth that are based on the so-called principle of charity; and thus (iii) a potential rehabilitation of the notion of disparate conceptual schemes—or, as it is better to say, of "constituted domains of objects".

The basic Kantian/Heideggerian conclusion can be summed up this way: the constituted objective world and the free constituting subject are intelligible only as two sides of one coin.

The Nature and Plausibility of Cognitivism

Cognitivism in psychology and philosophy is roughly the position that intelligent behavior can be explained (only) by appeal to internal "cognitive processes"—that is, rational thought in a 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.

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 1970/80, 1973/80), 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 counterintuitive 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, variable magnitudes.

The known regularities or relationships are expressed as equations (algebraic, vectorial, differential, or whatever) relating the values of various variables in any given situation: $F = ma = dp/dt$, for instance. 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 variables 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 variables, 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 macro-economics 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 (only) 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 that 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 nevertheless be able to follow the example.)

1 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;

2 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;

3 a projected image can be regarded as an array of closely packed dots of light, differing in brightness and color; and

4 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 perspicuity that it would gain. (Diagrams would help much more.) If we do not try to force it into a preconceived mold of scientific explanations, 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 1975b, 1973), 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, or 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 different use for it, and 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 is also morphological. The specified structure is little more than shape, and the specified abilities of what is so structured amount to rigidity, insolubility, and the like.

Now consider a case that 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 at stake. 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 so to 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. The 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,[1] but no one, to my knowledge, has previously distinguished morphological 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 (the doctrine 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 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 that 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 variables occurring in the equations of thermodynamic theory were found (or hypothesized) to correlate with quantities definable statistically in terms of the mechanical variables for groups of molecules. For example, the absolute temperature of a region was found to be proportional to the average kinetic energy of the molecules 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 in derivational-nomological; otherwise there would be no equational laws to derive, and probably no quantitative variables to occur 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 fairly complex and, for all I know, systematic.

In explaining a system, almost all 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, the distributor system of a car is a component in the (larger) ignition system, which, in turn, is a component in the complete engine system. Such a multilevel structure of nested systems is a *systematic hierarchy*. (See Simon 1969/81 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 traditional 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 would 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 independent, even though one is of the presuppositions of the other (compare Putnam 1973).

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 1974/81), nor are they worried if none are possible (*pace* Davidson 1970/80).

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 (1960) and the refinements it has inspired.[2] His 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 must now 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 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 is 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)

1 one (or another) particular chess move,

2 the solution of a logic problem, or

3 a scurrilous remark about the Queen of England and the Bishop of Canterbury?

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.[3]

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 (*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 typographic characters used in 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.

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

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 (for example, 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—the 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.

One important complication with this should be spelled out explicitly. Since the inputs and outputs "make sense" in virtue of a *pattern* they exhibit—a pattern that is extended in time—an IBB interpretation can also attribute enduring intentional "states" (and changes

therein) to an object. For instance, a blind chess player (person or machine) must keep track of, or "remember" the current position, updating it after each move; and any player will continuously "know" the rules, "desire" to win, and so on. In some cases, an input/output pattern can be so complicated that no sense can be made of it at all without attributing a rich (though slowly varying) "inner life" of beliefs and desires. It is important to realize, however, that this is "inner" only in the sense of an interpolation in the (external) input/output pattern—nothing is being said about the actual innards of the object.

Sometimes it will 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 example with which this section began is an IBB; so are adding machines, logic-problem solvers, automated disease diagnosticians, and (applying the definitions fairly 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 medical diagnostician it is giving good diagnoses relative to the symptoms provided. For reasons beyond the scope of this discussion, I don't think any effort to articulate completely general cogency conditions can succeed (though various authors have tried, at least for interpreting creatures as language users[5]). 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.[6] 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 …, and so on), 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 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 (for instance, 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 saw only how the notion of IBB could have empirical content, 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 be empirically defensible, 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" (that is, systematic) because rational 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 ideal 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 information processing system.[7]

This suggestion rests 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 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 those 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 program components are interpreted as operations on complex lists. So interpreted, these components 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.[8] Thus, the required abilities of bottom-level chess player 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 IPSs 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 instantiations, 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-/reinterpretation 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.

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 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 quantitative magnitudes 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 evidence of encapsulating what's relevant into a few highly specific interactions and abilities. By contrast, a resistor can be a functional component, because (almost) nothing about it matters except the way it resists the flow of electricity from one of its leads to the other. (Compare Simon 1969/81 on "partial decomposability"; and Marr 1977 on type-1 versus type-2 theories.)

So if neurons are to be functional components in a system, 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 just 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. Fairly detailed hypothetical models have been proposed for how holograms might be realized in neural structures; and there is some empirical evidence that some neurons behave in ways that would fit the models.[9]

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. First, they are prepared from the light bouncing off an ordinary object, and can subsequently be used to reconstruct a full three-dimensional 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 not 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.[10]

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.[11] 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. 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 some of what is known about how human chess players perceive the board and their options.[12] Implementation of such a device by optical means might well be impossible; but it is worth pointing out how much more general the neural medium (potentially) is. In the first place, transforms other than the Fourier transform could be implemented just as easily—including, perhaps, "custom" transforms for particular problems. Second, n-dimensional transforms are easily possible. Third, since neurons are connected "point-to-point", even the analog of an ordinary hologram wouldn't have to be arrayed as a surface—

physically, the "dots" could be distributed *ad libitum*, making possible all kinds of mingling and interaction among distinct "images". I have no clear idea what difference any of this would make; but it seems likely that the differences could be substantial. And, after all, the capabilities of regular holograms would have been difficult to visualize not so long ago.

Again, the point is that no "plausible-move generator" based on principles anything like this speculation would be an IPS. Nothing in it would "reason through" the move and counter-move alternatives that rationalize any move it proposed. Yet a chess player is a paradigm of what the "what-else-could-it-be" argument should apply to. (It's no accident that chess players are the most common IPS example.) 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.

7.1 Moods

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 versus felt qualities or the purely sensuous given.[13] 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.[14] 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.[15] 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. Finally, it may even be that some emotions (such as gratitude and regret) can be accommodated with a standard elaboration of this same segregation strategy—roughly, by treating them as compound states, with a cognitive (representational, propositional) component, and a separate noncognitive (qualitative, feeling) component.

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 grayer, 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 face 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.

7.2 *Skills*

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 (≠ 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 (1972/92, 106) asks an interesting pointed question about this ploy, in the special case of chess skills. 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)—that is, 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 noncognitivist 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.

7.3 UNDERSTANDING

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,[16] 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 theories, they grope for new insights into the phenomena, or new accounts that "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 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 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 constitute 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 other's 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 magnitudes 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.

It seems to me, however, that there is yet a deeper side to this: understanding pertains not primarily to symbols or rules for manipulating them, but to the world and 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 institutions. It is far from clear that these are governed by fully explicable rules at all. Our talk of them is sensible because we know 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, in their view, 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" quasilinguistic representations—only now they're sullied by omissions, probabilities, and heuristics. Deported from the immortal soul, however, they forfeit their original epistemic anchorage in the honesty of God and the natural light of reason. So, bereft of credentials from above,

the distinction of certain procedures as "reasonable" floats adrift, unless it can otherwise be explained. Evolution comes vaguely to mind, but much more needs to be said. My own hunch is that the intelligibility of rational "theorizing" 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" (Aristotle), even creativity and wisdom. We will understand understanding when we understand its many forms, primordial and refined. In the economy of understanding, words are merely money.

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 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 sub-discipline 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 mainly 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 it 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 morals 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 that 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. 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, 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 (1962/70) can be extended in a way that sheds light on a situation like this. 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 mean-

ingful 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 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 greatness 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 psychologically 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.

NOTES:

1 For instance, Cummins 1975; Putnam 1973; Dennett 1971/78; Simon 1969/81; and Fodor 1965

2 For instance, Davidson 1970/80, 1973/84; and Harman 1973; and compare Sellars 1954/63; Dennett 1971/78; and McCarthy 1979.

3 Compare Goodman 1968, chapter 4; Quine 1960, section 18.

4 Compare "structured description", Pylyshyn 1978.

5 For instance, Wilson 1959; Quine 1960, chapter 2; Lewis 1974/83; Grandy 1973; and Davidson 1973/84.

6 Compare McCarthy 1979 on thermostats.

7 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, chapter 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, as 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.

8 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—for example, not only how it manages to play chess, but also why it always neglects certain options.

9 For introductions to holograms and some of their interesting properties, see Leith and Upatnieks 1965; Herriott 1968; Gabor 1969; Firth 1972; and Cathey 1974. For further speculations about their capabilities, see van Heerden 1968; Pribram 1971, 1974; Pribram et al 1974; and Pollen and Taylor 1974. For hypothetical models of holograms realized in neurons, see Kabrisky 1966; Baron 1970; and Cavanagh 1972. For evidence that some neurons may actually behave in the required ways, see Campbell 1974; and Pollen and Taylor 1974; and compare Erickson 1974.

10 For some proposals of ips-based pattern recognizers, see Minsky and Papert 1972; and Waltz 1972.

11 See Yevick 1975, however, for a mathematician's tentative proposal of a holographically based logic.

12 De Groot 1965; Hearst 1967; Frey and Adesman 1976.

13 See Sellars 1956/63 for a discussion of this distinction in the context of a different issue.

14 Shoemaker 1975 and Block and Fodor 1972; but see also Dennett 1978a/78.

15 See, for instance, Shepard and Metzler 1971; Pylyshyn 1973, 1978; Paivio 1975; Kosslyn and Pomerantz 1977; Dennett 1978b/78.

16 See Winograd 1972, for example; and compare Greeno 1977.