Mind Design II

Philosophy
Psychology
Artificial Intelligence

Revised and enlarged edition

edited by
John Haugeland

A Bradford Book
The MIT Press
Cambridge, Massachusetts
London, England

While it may be true that normative discourse cannot be replaced without remainder by descriptive discourse, it would be a distortion to represent this as the aim of those who would naturalize epistemology. The aim is rather to enlighten our normative endeavors by reconstructing them within a more adequate conception of what cognitive activity consists in, and thus to free ourselves from the burden of factual misconceptions and tunnel vision. It is only the *autonomy* of epistemology that must be denied.

Autonomy must be denied because normative issues are never independent of factual matters. This is easily seen for our judgments of instrumental value, as these always depend on factual premises about causal sufficiencies and dependencies. But it is also true of our most basic normative concepts and our judgments of intrinsic value, for these have factual presuppositions as well. We speak of *justification*, but we think of it as a feature of *belief*, and whether or not there are any beliefs and what properties they have is a robustly factual matter. We speak of *rationality*, but we think of it as a feature of *thinkers*, and it is a substantive factual matter what thinkers are and what cognitive kinematics they harbor. Normative concepts and normative convictions are thus always hostage to some background factual presuppositions, and these can always prove to be superficial, confused, or just plain wrong. If they are, then we may have to rethink whatever normative framework has been erected upon them. The lesson of the preceding pages is that the time for this has already come.

Connectionism and Cognition

11

Jay F. Rosenberg
1990

I propose to preach a modest sermon against the mediaeval sin of *Enthusiasm*. There's a bright and powerful new paradigm abroad in the philosophy and psychology of mind—the *connectionist* paradigm of brainlike neural networks, distributed representations, and learning by the back propagation of error—and I fear that it is well on the way to becoming a new gospel as well. It has its array of saints and prophets—McClelland and Rumelhart, Hinton and Sejnowski, Churchland and Churchland—and it has its ritual observances—spirited meetings of the San Diego Traveling Connectionist Extravaganza, complete with tape recordings, video cassettes, and multicolored overhead transparencies. It is a very impressive business indeed.

There can be no doubt that the connectionist paradigm has equipped us with potent new tools for understanding *something*, even *many* things. It is less clear, however, just *what* the connectionist paradigm equips us to understand. The San Diego Enthusiasts tend to say "everything mental"—that's what makes them capital-E Enthusiasts—but especially they tend to say "cognition", and at least one of them has begun to talk of the connectionist paradigm as the opening wedge of a global challenge to "sentential epistemologies" in general. Here, however, I do have my doubts, and these doubts are what I want primarily to talk about. But first, even the devil must be given his due—and connectionism is certainly no devil.

What *can* brain-like connectionist networks (of the sorts that I trust I can assume are familiar) help us to understand? I think there are at least three impressive accomplishments, each well worthy of being welcomed with some (small-e) enthusiasm.

First, connectionist networks give us considerable insight into the specific mechanisms by means of which the brain might function as a *transducer*. This role has a variety of aspects. NETtalk (Sejnowski and Rosenberg 1987), for example, neatly illustrates one of them, the

transposition of sensory modalities, in this instance from visual inputs into acoustic outputs, printed words into audible speech. More broadly, as Paul Churchland (1986) has elegantly argued, connectionist networks allow us to begin to address the problem of *motor control*, for instance, eye-hand coordination. The cerebellum, in particular, appears to be structured in layers whose intra- and interconnectivities echo those of a multilayered connectionist network. The partitionings of the corresponding hidden-unit activation-vector phase spaces can then be understood as "maps" of the organism's visual and tactile environments whose interconnections (forming a "phase-space sandwich") constitute, from the formal point of view, a "matrix multiplier" that can directly instantiate the sophisticated mathematical transformations requisite for smooth sensory-motor coupling.

Second, the connectionist paradigm equips us with the valuable notion of a *distributed representation*. It exhibits a method of *globally* encoding a family of discriminations—in a configuration of connection weights that induces a family of partitions on a connectionist system's hidden-unit vector-activation phase space interpretable, in turn, as a structure of prototypes and associated similarity metrics—which allows us to address problems classically formulated under the rubrics of *pattern recognition* and *information retrieval* (without positing exponential searches), and to understand the *graceful degradation* of information-carrying systems (without positing massive reduplications).

Third, supplemented by the mechanism of back propagation of error, the connectionist paradigm gives us a handle on certain traditional problems of *learning*, and the possibility of a system's acquiring discrimination capacities ab initio. Such examples as the training-up of a connectionist network to respond differentially to sonar echoes from rocks and from mines (Gorman and Sejnowski 1988) shows us vividly how an initially unstructured back-propagation system can function as a *pattern-extractor* given only a suitably rich set of inputs.

Those are significant accomplishments indeed, and they take us a good distance toward understanding how some of the achievements traditionally characterized as "mental" might be operationally realized in organs structured in much the way the human brain is evidently structured. It is not surprising, on that account, to find more than one Enthusiast arriving promptly at the conclusion that what we need to understand the residue of human "mentality" is, in essence, more of the same.

> If even small artificial networks can perform [such] sophisticated cognitive tasks ..., there is no mystery that real networks should do the same or better. What the brain displays in the way of hardware is not radically different from what the models contain, and the differences invite exploration rather than disappointment. The brain is of course very much larger and denser than the models so far constructed. ... It plainly commands many spaces of stunning complexity, and many skills in consequence. It stands as a glowing invitation to make our humble models yet more and more realistic, in hopes of unlocking the many secrets remaining. (p. 187 [283])

Thus says Paul Churchland, in his essay "On the Nature of Theories: A Neurocomputational Perspective" (1990/89 [chapter 10 in this volume]), which will be my chief stalking horse here. Setting aside for the moment the question of whether these accomplishments are properly described as "sophisticated cognitive tasks", the first thing that needs to be said about such remarks is that, even *if* they are properly so described, the only sense in which small artificial connectionist networks *perform* them is the sense in which my portable computer regularly performs even *more* "sophisticated cognitive tasks": computing amortization tables, correcting misspellings in my documents, and roundly defeating me at games of Reversi.

The habit of thus nonchalantly importing the personal vocabulary appropriate to what Dennett calls "intentional systems" (1971/78, 1981/87 [chapter 3 in this volume])—prototypically human beings—into descriptions of the operations and functions of *subpersonal* systems (brains, for example) is one of the sure signs of Enthusiasm, and Churchland is one of the worst offenders in this regard. Here he is discussing the rock/mine network:

> [During the training period,] the system is *theorizing* at the level of the hidden units, *exploring* the space of possible activation vectors, *in hopes of finding* some partition or set of partitions on it that the output layer can then *exploit* in turn, so as to *draw the needed distinctions* and thus bring the process of error-induced synaptic adjustments to an end.
>
> (pp. 179f [278]; all but the first emphasis mine.)

This transposition of a characteristically personal vocabulary to subpersonal systems is, I think, an essential element of Churchland's strategy in proposing that the new connectionist paradigm finally supplies a "comparably compelling *alternative* conception of representation and computation" to the "sentential epistemologies" whose poverty he has

been urging for more than a decade. Such "sentential epistemologies", he says (p. 154 [252]), are characterized by two fundamental assumptions:

(1) that language-like structures of some kind constitute the basic or most important form of representation in cognitive creatures, and

(2) that cognition consists in the manipulation of those representations by means of structure-sensitive rules.

It is not entirely clear, however, what either of these claims amounts to. To begin with, in the case of (1), there are surely *many* senses in which a structure might properly be described as "language-like" and in which language-like structures might constitute the "basic" or "most important" form of representation in cognitive creatures. In particular, a structure, or, better, a family of structures, could be "language-like" in the strong sense of being usefully characterized as having *logical form*, and thereby as instantiating a *compositional syntax* strongly analogous to the linear, concatenative, recursive syntax of a formal system or a spoken or written natural language. In a much less restrictive sense, however, a family of structures might be "language-like" only functionally, in being usefully characterized as having *propositional form*, that is, as representing states of affairs by both *referring* to objects and *characterizing* them as being such-and-such. In the latter sense, graphs and pictographs, hieroglyphics, ideographs, portraits, photographs, and maps could all, in different ways, qualify as "language-like structures"—and so too, perhaps, even the distributed representations encoded in some trained connectionist networks.[1]

Again, to sound an Aristotelian note, language-like structures (of some determinate sort) might well turn out to be "basic" or "most important" in one respect, say in the order of knowing, without being "basic" or "most important" in another, such as the order of being. Something like this would be the case, for example, if, as has been proposed (Bechtel 1988a, 1988b), connectionist networks stand to some traditional "rule-following" information processing systems as microstructure to macrostructure, an underlying framework in terms of which those traditional systems are *implemented*. Again, the difference between the connectionist and traditional models of a single information-processing system might turn out to be a function primarily of the level of analytical regard. A fully-trained NETtalk system, for example, can be viewed "syntactically", simply as an interconnected

system of activation-weighted nodes, but *also* "semantically", as encoding both the 79 fundamental letter-to-phoneme correlations requisite for transposing written into spoken English (in its partitioning of its hidden-unit activation-vector phase space) and the hierarchical organization of the phonetic structure of English speech (straightforwardly recoverable by a cluster analysis of that partitioning).

Similarly, the assumption, formulated in (2), that cognition consists in the "manipulation" of such language-like representations "by means of structure-sensitive rules" admits of various understandings. These range from the strong, but wildly implausible, view that cognitive activity is a species of (deliberate, self-conscious) *rule-obeying* conduct—a game in which explicitly formulated (meta-level) representations of rules function as reasons, authorizing transformational "moves" from one representation to another—to the much less striking, but almost inescapable, view that cognitive activity is (at least) a species of *rule-conforming conduct*—a family of practices that (at least) accord with permissible "moves" of some explicitly formulable representation-transformation game.

The word 'cognition' is, of course, nobody's personal property, but unless the intent is to weaken the term's commitments beyond all recognition, one thing that should be clear, at least since the publication of Sellars's "Empiricism and the Philosophy of Mind" (1956/63), is that the mere exercise of a discrimination capacity, however complex, is not yet an example of *cognition*. A magnet quite efficiently discriminates between ferrous and nonferrous materials, but that does not put it in the running for the title of "cognitive system". Just as one can "train up" a modest connectionist network regularly to respond differentially to sonar echoes from mines and those from rocks—or, more precisely, as it turns out, to sonar targets made of metal and those made of nonmetal—I can (by stroking it with a strong magnet) "train up" a screwdriver regularly to respond differentially to brass and steel screws. There is no more reason to regard the trained network's acquired response to a (metal) mine as its possession of an ur-concept of metal (or ur-*awareness* of the mine as made of metal) than there is to ascribe an ur-*concept* of steel or (ur-awareness of certain screws as made of steel) to the "trained" (magnetized) screwdriver simply on the basis of its acquired propensity to respond differently to steel and brass screws.

Churchland apparently would not quarrel with these last remarks. Indeed, he himself writes that:

> It is briefly tempting to suggest that NETtalk has the concept of 'hard *c*', for example, and that the rock/mine network has the concept of 'metal'. But this won't really do, since the vector-space representations at issue do not play a conceptual or computational role remotely rich enough to merit their assimilation to specifically human concepts. (pp. 175f [274])

But if this is right, as it surely is, then Churchland's later sanguine characterizations of the accomplishments of such connectionist systems as NETtalk and the rock/mine network as the performance of "sophisticated cognitive tasks" is just so much Enthusiastic hyperbole. Whatever else performing a "sophisticated cognitive task" requires, it at least requires some sort of utilization of concepts. It follows that, if a connectionist system's arriving at a determinate stable partitioning of its hidden-unit activation-vector phase space does not count as its possessing or having mastered a concept, then neither will its ensuing successful discriminations count as *cognitive* performances, sophisticated or unsophisticated.

At this point, Churchland might well object that he has not rejected the identification of connectionist vector-space representations with concepts in general, but only the assimilation of such representations to "specifically human" concepts. Thus, while it would not be correct to say that the trained rock/mine network has the concept of 'metal' *as opposed to*, for example, the concept 'mine' (or some other concept extensionally equivalent over the training class of inputs), the fact that the partitioning of its vector-space is not only stable but also *generalizable*, in that the network successfully classifies *new* sonar echoes from both rocks and mines, warrants our ascribing to it primitive or rudimentary concepts at least of two *kinds* of sonar targets. *We*, given further experimentation and our more sophisticated representational resources, can then subsequently come to recognize and identify these as primitive concepts of metal and nonmetal as opposed to (the locally extensionally equivalent distinction between) rocks and mines.

Now I do not know whether Churchland would in fact adopt the conciliatory strategy I have just been outlining, but in any event, I want to resist the particular sort of blurring of useful distinctions I am convinced it represents. Although I have already granted that no one *owns* the term 'concept', in contrast to Quine's (1948/53) cheerful generosity to McX vis-à-vis the word 'exists', I am reluctant just to give Churchland the word 'concept' and go off in search of an alternative idiom for my own use. Instead, I would like to hold fast to the Kantian

insight that the notion of a judgment is *prior* to that of a concept—"the only use that the understanding can make of ... concepts is to judge by means of them" (1787/1929, A68=B93)—and that, consequently, since whatever else a *judgment* may be, it is something fitted to play the role of a premiss or conclusion in *reasoning*, there is an essential connection between the notion of a concept and that of *inference*. Sellars (1981) analogously argues that we must be careful not to conclude straightaway that a rat which has acquired a propensity to leap at panels with varieties of triangles painted on them has, simply by virtue of its training, acquired an ur-concept of a triangle:

> To suppose that it *has* reflects the common conviction that the connection between representational states and objects is a direct one-one correlation. Obviously, the representational state ("symbol") is correlated with what it represents—but this correlation may *essentially* involve other correlations—thus between it and other representational states and between representational states and action. (p. 335)

What differentiates the exercise of a mere discriminative capacity, a systematic propensity to respond differentially to systematically different stimuli, from a conceptual representing properly so called, is that the latter has a place and a role in a system of *inferential* transformations, a web of consequences and contrarieties in terms of which the representational state admits of being (more or less determinately) located in a "logical space" with respect to *other* representations.[2]

> [A particular state of a rat, for example] wouldn't be a state of representing something as a triangle, unless [the rat] had the propensity to move from [that state] to another state which counts as a primitive form of representing it as three-sided or as having, say, pointed edges. (p. 336)

Churchland, we recall, grants that the vector-space representations generated by NETtalk or the rock/mine network "do not play a conceptual or computational role remotely rich enough to merit their assimilation to specifically human concepts" (p. 177 [274]), but neither does he pause to tell us what sort of "conceptual or computational role" would be "rich enough". One might suppose, then, that he could and would accept with equanimity our most recent remarks connecting the notion of specifically *conceptual* content of representations to their inferential roles. This, however, would be to misread him. For consider how the passage from which I have just quoted continues:

> Nevertheless, it is plain that both networks have contrived a system of internal representations that truly corresponds to important distinctions and structures in the outside world, structures that are not explicitly represented in the corpus of their sensory inputs. The value of those representations is that they and only they allow the networks to "make sense" of their variegated and often noisy input corpus in the sense that they and only they allow the network to respond to those inputs in a fashion that systematically reduces the error messages to a trickle. These, I need hardly remind, are the functions typically ascribed to *theories*.
>
> What we are confronting here is a possible conception of knowledge or understanding that owes nothing to the sentential categories of current common sense. A global theory, we might venture, is a specific point in a creature's synaptic weight space. It is a configuration of connection weights, a configuration that partitions the system's activation-vector space(s) into useful divisions and subdivisions relative to the inputs typically fed the system. 'Useful' here means: tends to minimize the error messages.

The problem is that an account of a "rich enough conceptual or computational role" for a connectionist network's representations which takes as its model *inferential relations among propositions* could hardly be said to "owe nothing to the sentential categories of current common sense". On the contrary, such an account would retain precisely what is *essential* to traditional "sentential epistemologies"—namely, items that have propositional form and stand in logical relations, while sloughing off as *adventitious* only the fact that, in natural languages, the referring and characterizing *functions* requisite for propositional form are characteristically performed with the aid of distinct notational devices (utterance tokens or sign-designs).[3]

Two further aspects of the passage we have just been examining deserve some comment. Of course the application of such personal psychological predicates as "contrives" and "make sense" (with only the latter in cautionary quotes) to these small and distinctly subpersonal networks is once again best understood as an outcropping of Enthusiastic excess. But so too, and more significantly, I want to suggest, is the not-so-innocent definite article in Churchland's phrase "*the* functions typically ascribed to theories". For while classifying or redescribing the elements of a corpus of sensory inputs in terms positing distinctions and structures in the world that are not explicitly (phenomenally) represented in those inputs is certainly *a* function of theories, only someone intent on constructing an extraordinarily impoverished view of natural science could possibly speak of it as *the* function of theories.

Now Churchland is notoriously not an advocate of an impoverished view of natural science. He is a scientific realist, who has no patience with fictionalist or instrumentalist views that assign any special *epistemically* privileged status to the observational concepts of common sense. But, for all that, I do not think that we can write off the definite article here as a mere *lapsus linguae*. As Churchland very well knows, a theoretical redescription of some family of observable phenomena is essentially a prolegomenon to the theoretical *explanation* of those phenomena as phenomena. It is the explanatory subsumption of the redescribed phenomenon under *laws*, belonging to a system of *inferentially interrelated* principles, that gives the redescription its point. But that, of course, just is "sentential epistemology" all over again. The problem, however, is that the *only* thing a connectionist network ever learns to do is to partition its input. Since Churchland (like everyone else) has no idea how to pry the notion of an explanatory understanding of phenomena as phenomena loose from traditional sentential epistemics, his only choice, in order to characterize such networks as *theorizers*, is subtly to scale down the notion of a *theory* until it fits their limited competences.

The second aspect of the quoted passage worth attending to is that it highlights again the role of the notion of *error* in the connectionist paradigm. Now, in the course of examining the question of how faithfully connectionist networks depict the organic brain, Churchland does raise some questions about the *de facto applicability* of the connectionist model of learning by the back propagation of apprehended errors to real biological systems. What he does not do, however, is to raise any questions regarding the *sense* of the notion of "learning by back propagation of apprehended errors" in its envisioned application to natural creatures. Specifically, he does not pause to inquire in virtue of *what* some particular output from a connectionist network could be apprehended *by that network* as an *error*.

What makes this question worth asking is that it is precisely at this point in the connectionist story that its paradigm acquires whatever *normative epistemological* import it has. Churchland claims that we now possess "a powerful and fertile framework with which to address problems of cognition ... that owes nothing to the sentential paradigm of the classical view" (154 [252]), and he professes to be deeply skeptical about even the very notion of truth:

> It is no longer clear that there is any unique and unitary relation that virtuous belief systems must bear to the nonlinguistic world. Which leaves us free to reconsider the great many different dimensions of epistemic and pragmatic virtue that a cognitive system can display. (p. 157 [255])

But when the chips are down, what drives the connectionist picture of learning is an *assumed* bipolarity of "correct" versus "erroneous" responses—and since distributed connectionist representations *can* be functionally understood as "language-like"—that is, as having propositional form—this bipolarity is close enough to "true" versus "false" beliefs as to make no difference.

Churchland is not, of course, insensitive to these points, and, indeed, remarks on them in the course of (pessimistically) assessing the plausibility of the idea that the back propagation of apprehended error is in fact the central mechanism for learning in organic brains.

> A necessary element in [the delta rule's] calculated apportionment of error is a representation of what would have been the *correct* vector in the output layer. This is why back propagation is said to involve a global *teacher*, an information source that always knows the correct answers and can therefore provide a perfect measure of output error. Real creatures generally lack any such perfect information. They must struggle along in the absence of any sure compass toward the truth, and their synaptic weights must undergo change, change steered in some way by error or related dissatisfaction, change that carves a path toward a regime of decreased error. (p. 186 [282])

Now the brains, of course, do *not* learn; the creatures do. But to the extent (evidently considerable) that a creature's brain is or resembles a system of networks of the connectionist sort, the configurations of its synaptic weights surely must undergo changes as the creature learns. One can sensibly say that such changes are "steered in some way by error" and "carve a path toward a regime of decreased error", however, only if one is *either* equipped with an *antecedent* notion of "correct representation" or prepared to say that a "correct representation" just *is* whatever configuration of weights *in fact* stabilizes the system over the actual inputs to it.

Churchland makes it unmistakably clear that he would be loathe to adopt the latter course, along with its "convergent realist" implication that we cannot but arrive at correct representations of the world.

> For one thing, nothing guarantees that we humans will avoid getting permanently stuck in some very deep but relatively local error minimum. For another, nothing guarantees that there exists a possible configuration of weights that would reduce the error messages to *zero*. (p. 194 [289])

Such remarks make sense, however, only if the notion of an "error"—or equivalently, that of a "correct representation"—admits of characterization independently of the *de facto* achievements of connectionist representers. One would expect, then, to find Churchland offering us an account of the *antecedent* notion of "correct representation" required for speaking sensibly of "error" in this connection in the first place. But one does not. What one finds instead is another subtle Enthusiastic blur.

That a connectionist system learns by the back propagation of *error* requires that its hidden-layer activation weights be adjusted in the direction of decreasing the difference between what its outputs *are* and what those outputs *ought to be*. It is only this implicit appeal to a *norm*-driven model of learning, I suggest, that makes it appropriate to describe the activities and accomplishments of such systems in *epistemic* terms at all. At the crucial juncture, however, Churchland speaks, not simply of "error", but more cagily of "error or related dissatisfaction". Now we can certainly imagine highly plastic "self-teaching" organic connectionist systems which are hard-wired to "learn" by adjusting their activation weights so as to minimize *something*—perhaps, indeed, something describable as a "dissatisfaction", such as pain, hunger, thirst, fatigue, or what have you. But to characterize the resultant accomplishments (for instance, acquired discrimination capacities, and inter-sensory or sensory-motor coordinations) in *epistemic* terms is an Enthusiastic vice of the sort against which Sellars cautioned us over thirty years ago.

> [The] idea that epistemic facts can be analyzed without remainder—even "in principle"—into non-epistemic facts, whether phenomenal or behavioral, public or private, with no matter how lavish a sprinkling of subjunctives and hypotheticals is ... a radical mistake—a mistake of a piece with the so-called "naturalistic fallacy" in ethics. (1956/63, p. 257/131)

The essential point is that characterizing a state or transaction or condition or episode in epistemic terms is not providing an empirical, mater-of-factual description of it, but rather locating it within a

"logical space" of *reasons* and *justification*. A creature properly in this logical space of justification is one capable of recognizing or acknowledging the superior (epistemic) authority of some representings vis-à-vis others. Such a creature responds to epistemic authority, for example, by adopting or endorsing some (implied) representations *because* they are consequences of others and by abandoning or modifying some (contrary) representations *because* they conflict with others to which it is committed. This, not surprisingly, brings us around again, although at a deeper level, to the notion of inference that I have already argued is essential to distinguishing the exercise of mere discriminative capacities, however sophisticated, from authentically cognitive performances, however primitive.

What needs to be stressed is that, for a creature properly to be said to move in the logical space of reasons and justification, it is not enough that it be usefully characterizable as "rational", in the sense, for example, of behaving in ways fruitfully described, understood, and predicted from Dennett's intentional stance. Such a creature, we may say, is "logic-*conforming*"; but a creature capable of acting for reasons *acknowledged as* (epistemically) authoritative and of responding to errors apprehended *as* errors must do more than merely behave in ways that conform to logic. It must *use* logic.

The distinction between merely logic-conforming or rational creatures and logic-using or *ratiocinative* creatures, although clear enough in the abstract, is one that remains unmarked by the intentional stance as such. From the intentional stance, the behavior of a deer who flees when it scents smoke can perhaps fruitfully be explained by ascribing to the deer a belief that there is a fire nearby and a desire to avoid perishing in it. The deer, we may say in a Humean tone of voice, has learned to "associate" smoke with fire, and it is this "association" that, when smoke is present, gives rise to its representation of a belief. Furthermore, since the presence of smoke is indeed evidence that a fire is nearby, we may go on to say that the deer then "has a good reason" for believing that a fire is nearby. Its belief is "well grounded".

The Enthusiastic mistake which needs to be avoided here does *not* lie in this move from "the deer represents the presence of smoke, and the presence of smoke is a good reason for believing that a fire is nearby" to "the deer has a good reason for believing that a fire is nearby", but in the misinterpretation of this line of thought as licensing the further conclusion that the deer acknowledges and responds to the reason that it "has" *as* a reason—in other words, to the available

evidence *as evidence*. This stronger claim requires that the deer be capable of *representing* its evidence *as* evidence. That is, not only must the deer have the propensity to represent that a fire is nearby whenever it represents that smoke is present (a propensity it shares with the smoke detector in my apartment), but it must also be equipped, so to speak, in *some* way, to form a judgment to the effect that the presence of smoke is evidence that a fire is nearby.

This is not yet to demand that the deer be in possession of generic epistemic concepts as such—that is, *concepts* of evidence, good or poor reasons, correct or erroneous representations, and so on—although, obviously, any such *epistemic* creature would satisfy the condition in question. There is an intermediate sense in which even a creature rather like our deer could be said to respond to the reason it "has" *as* a reason, or to its evidence *as* evidence. What is needed is that its representation of the "conclusion" that there is a fire nearby be mediated by a representation of the *material counterpart* of the specific epistemic relation of "good evidence" in question, namely, by a representation of a *generalization* to the effect that whenever smoke is present anywhere, there is likely to be a fire nearby there. But this would imply, in turn, that the creature in question must be not only a rational or logic-conforming creature, in the sense we have recently been exploring, but also something that no actual deer is, a ratiocinative or logic-using creature as well.

As Leibniz put it, in passing from a representation of smoke as present to a representation of fire as nearby *simply* in accordance with an acquired propensity to do so, the deer does not reason but manifests only "a sort of *consecutiveness* which imitates reason".[4] The crucial point for our present purposes is that genuinely ratiocinative creatures must be capable of a range of representations that extends significantly beyond those that the connectionist paradigm admittedly helps us to understand. Ratiocinative creatures, that is, must be capable not only of representings that have propositional form and thereby stand in logical relations of consequence and contrariety, but also of representings of representings *as* thus logically related, and consequently of those logical relations themselves. In Sellars's words:

> [We] carve nature at the joint by distinguishing between those [representational systems] which contain representational items which function as do logical connectives and quantifiers, that is, which have logical expressions "in their vocabularies", and those which do not.
>
> (1981, p. 341)

What these considerations help us recover is a second, more robust, sense of 'rational' in which a "rational creature" is one not only fruitfully understood as manifesting the sort of "practical rationality" definitive of the intentional stance but also meaningfully characterizable as possessing "theoretical rationality" as well.[5] This comes about as follows.

A ratiocinative creature is capable not only of generalizable representations, but also of representations of generalities. Again, while a system at the level of complexity of, say, the rock/mine network might, with some license, be said *not to* "believe" of a given input that it is the echo of a mine, only a genuinely ratiocinative system, possessing resources adequate for representing negation, could sensibly be said to "believe" of a given input *that it is not* the echo of a mine. Such resources, however, enable a ratiocinative system to do what a connectionist system of the sort we have been examining cannot do, namely, to "internalize" a test for erroneous representation in the form of a global constraint of logical consistency. It shows us, that is, how a representation could in principle come to be apprehended as an error *by the system itself.*

Two representations having the *logical* forms "All A's are B's" and "This is an A, but not a B" cannot both be correct; and while such an inconsistency by itself is, of course, insufficient to determine which of the representations is in error, in a ratiocinative system operating under a global consistency constraint, it can certainly suffice to set into motion precisely the sort of homeostatic "web of belief" interadjustments of representational and inferential commitments that are the stock in trade of *sentential* epistemologies. There is a clear sense, therefore, in which a ratiocinative system not only responds to reasons *as* reasons, but, even without the aid of a "teacher", can respond to errors *as* errors as well. (The role of global consistency constraints of this sort is provocatively discussed in Millikan 1984.)

The moral to be carried away from this discussion is not that a ratiocinative system cannot in principle be thought of as assembled from connectionist-style resources. Although we do not have the slightest idea how representations having the logical forms of conditionals and negations might be encoded in the "distributed" way appropriate to connectionist networks, none of the considerations we have adduced implies that they *cannot* be so encoded. And, *if* they can, then there is also no reason to reject out of hand the suggestion that the homeostatic "web of belief" interadjustments among such representational

and inferential commitments take the de facto operational form of adjustments of the activation weights of the hidden units of some complex multi-layered system of the connectionist sort.

The point, however, is that *this* connectionist story is *essentially* the story, not of an alternative to a sentential epistemology, but of the *implementation* of a sentential epistemology. The "language-like" character of the logically articulated representations thus encoded is not adventitious, but necessary for us to be able to understand the operation of such a system in *epistemic* terms at all. Absent even the minimalist interpretations of "responding to reasons as reasons" and "responding to errors as errors" that first become possible in the case of a system understood as ratiocinative—given, for example, a system exhibiting only the "dissatisfaction minimizing" learning envisioned in passing by Churchland—the specifically epistemic vocabulary finds no point of purchase.

The sin of Connectionist Enthusiasm, we have seen, takes many forms. It expresses itself in sanguine applications of the personal epistemic vocabulary to subpersonal systems. It expresses itself in the subtle paring down of rich epistemic notions—in neglecting the inferential embeddings that minimally distinguish *concepts* from mere discrimination capacities and the explanatory applications that minimally distinguish *theories* from mere taxonomies. And it expresses itself in the blurring of distinctions among the diverse ways, structural and functional, in which a representational system can be "language-like".

Most significantly, however, Connectionist Enthusiasm manifests itself in its willingness to take *normative* epistemic force for granted. For the connectionist paradigm locates this normative epistemic force *outside* the network itself, in its "teacher's" infallible knowledge of which of its outputs are erroneous and of how they differ from the correct outputs—that is, from what they *ought* to be.

The resulting problem can be solved. Normative epistemic force can be "internalized" by a representational system which can respond to reasons *as* reasons and errors *as* errors. But it can be solved only by "splitting" the intentional stance, by distinguishing the ratiocinative from the *merely* rational, and so only by acknowledging representational systems whose representations are "language-like" in the strong sense of possessing not merely propositional form but logical form as well. Sentential epistemology, in short, is the only *epistemology* we've got, and connectionism is at best its implementationist underlaborer.

To say this is not to disparage the brilliance of the connectionist achievement. For what it shows us is nothing less than how we might begin to put sentential epistemology together with the organic brain, and that is well worth celebrating. As is so often the case on such occasions, however, some of the revelers have a tendency to celebrate to excess. They claim that what connectionism shows us is an epistemology, that is not sentential, and that the organic brain might use instead. But that, I have argued, is just Enthusiasm—and in philosophy, too, Enthusiasm is still a sin.

Notes

1. See, in this connection, Sellars 1981, to which the present discussion is, and will continue to be, deeply indebted.
2. The priority of the notion of a judgment to that of a concept is also, of course, a central principle of Frege's philosophy. Frege's strategy, recall, is precisely to replace a "bottom up" account of judgments in terms of the composition of concepts by a "top down" analysis of the notion of "conceptual content" in terms of intersubstitutivity of and in judgments *salva* correct inferences.
3. This fact, in turn, is explained by the (temporal) *linearity* of speech (and the spatial linearity of script). As Wittgenstein was the first to see—in the *Tractatus* (1922/74)—the predicate expressions of such a linear representational system are, from the *functional* point of view, auxiliary signs, serving only to guarantee a stock of characteristics of and relations among referring expressions [*names*] (that is, "being concatenated with a 'red'", "standing respectively to the left and right of a 'taller than'") adequate for representing possible characteristics of and relations among the objects to which those expressions refer.
4. *Monadology*, #26 (Leibniz 1714/1977), cited in Sellars 1981, p. 342.
5. It is this sense of 'rational', I think, that Jonathan Bennett proposes to isolate and examine in his delightful and insightful little book, *Rationality* (1964).

Connectionism and Cognitive Architecture: A Critical Analysis

12

Jerry A. Fodor
Zenon W. Pylyshyn
1988

1 Introduction

Connectionist or *PDP* models are catching on. There are conferences and new books nearly every day, and the popular science press hails this new wave of theorizing as a breakthrough in understanding the mind. There are also, inevitably, descriptions of the emergence of connectionism as a Kuhnian "paradigm shift". (See Schneider 1987, for an example of this and for further evidence of the tendency to view connectionism as the "new wave" of cognitive science.) The fan club includes the most unlikely collection of people. Almost everyone who is discontent with contemporary cognitive psychology and current "information processing" models of the mind has rushed to embrace "the connectionist alternative".

When taken as a way of modeling *cognitive architecture*, connectionism really does represent an approach that is quite different from that of the classical cognitive science that it seeks to replace. Classical models of the mind were derived from the structure of Turing and Von Neumann machines. They are not, of course, committed to the details of these machines as exemplified in Turing's original formulation or in typical commercial computers—only to the basic idea that the kind of computing that is relevant to understanding cognition involves operations on symbols (see Newell 1980, 1982; Fodor 1976, 1987; and Pylyshyn 1980, 1984). In contrast, connectionists propose to design systems that can exhibit intelligent behavior without storing, retrieving, or otherwise operating on structured symbolic expressions. The style of processing carried out in such models is thus strikingly unlike what goes on when conventional machines are computing some function.