METAPHILOSOPHY
Vol. 28, Nos. 1&2, January/April 1997
0026–1068

CONNECTIONISM AND THE PHILOSOPHICAL FOUNDATIONS OF COGNITIVE SCIENCE

TERENCE HORGAN

ABSTRACT: This is an overview of recent philosophical discussion about connectionism and the foundations of cognitive science. Connectionist modeling in cognitive science is described. Three broad conceptions of the mind are characterized, and their comparative strengths and weaknesses are discussed: (1) the classical computational conception in cognitive science; (2) a popular foundational interpretation of connectionism that John Tienson and I call "nonsentential computationalism"; and (3) an alternative interpretation of connectionism we call "dynamical cognition." Also discussed are two recent philosophical attempts to enlist connectionism in defense of eliminativism about folk psychology.

Key words: Connectionism; Connectionist Network; Cognition; Cognitive Science; Computation; Dynamics; Dynamical System; Eliminativism; Folk Psychology; Foundations of Cognitive Science; Language of Thought; Mind; Neural Network.

Connectionism is an important recent movement within the inter-disciplinary field known as cognitive science. The rise of connectionism has prompted much recent philosophical discussion. This paper is an overview of recent debates about connectionism and the foundations of cognitive science.[1]

Let me mention two preliminary points, before proceeding. First, I make no pretense that this survey is impartial. My colleague John Tienson and I are actively involved in current debates about connectionism, and we have staked out a fairly distinctive position on the issues. What I offer here is an overview of the landscape as seen from our own philosophical perspective.

[1] Other articles and books that are useful for familiarizing oneself with connectionism and with philosophical discussions of it include Bechtel (1991), Bechtel (1993), Bechtel and Abrahamsen (1991), Churchland (1988) pp. 146–55, Churchland (1995), Clark (1993), McLaughlin (1993), and Tienson (1988). Collections of papers about philosophy and connectionism include Horgan and Tienson (1991); Ramsey, Stich, and Rumelhart (1991); and *Synthese* 101, 3 (1994), Issue on Connectionism and the Frontiers of Artificial Intelligence, ed. A. Clark.

Second, I will say next to nothing here about one key issue in the philosophy of mind and in the foundations of cognitive science: the nature of mental intentionality. In both traditional and connectionist cognitive modeling, intentional content is assigned to certain states of a model, essentially by the modeler's fiat – aiming, of course, to respect the causal properties of both the model and the cognitive system being modeled. Models are described in terms of representations, and cognitive processing is construed as the system's evolution from one representational state to another. It is assumed that natural cognitive systems have intentional states with content that is not derived from the interpreting activity of other intentional agents. But it is not the business of either traditional or connectionist cognitive science to say where underived intentionality "comes from" (although each may place certain constraints on an answer to this question). Thus, intentionality as such is not at issue in the differing positions I will discuss, and will be mentioned only as it comes up in relation to other issues.

1. Classical Computational Cognitive Science (*Classicism*): A "Just-So Story" About the Mind

Connectionism arose as a rival to classical, artificial-intelligence style, cognitive science (*classicism*, as I will call it henceforth). Philosophical positions concerning connectionism have been largely articulated in terms of its perceived relations to, and differences from, classicism. So let me begin by discussing classicism.

David Marr (1982) provided an influential account of levels of description in computational cognitive science. His account brings into focus some of the most basic, foundational assumptions of classicism. Marr argued that in order to understand complex information-processing systems, one needs to consider them from three distinct theoretical levels. He wrote:

> At one extreme, the top level, is the abstract computational theory of the device, in which the performance of the device is characterized as a mapping from one kind of information to another, the abstract properties of this mapping are defined precisely, and its appropriateness and adequacy for the task are demonstrated. In the center is the choice of representation for the input and output and algorithm to transform one into the other. At the other extreme are the details of how the algorithm and representation are realized physically – the detailed computer architecture, so to speak. (1982, 24–25)

One can call these three levels (1) the level of the cognitive function,[2]

[2] Marr labels the top level 'the theory of the computation'. This is ambiguous, between *what* is to be computed and *how* it is to be computed. Clearly what Marr has in

(2) the level of representation and algorithm, and (3) the level of implementation or realization.

An algorithm, or program, is a mathematical object, a set of rules for manipulating symbols or data-structures purely on the basis of their formal/structural properties, independent of any intentional content they might have. Symbols and data-structures, so described, are also mathematical objects. Thus, the middle level in Marr's typology is a mathematical level of organization.

In this we have a central and important idea about cognitive design, which cognitive science has embodied from its inception: what's important about the brain, *vis-a-vis* mentality, is not its specific neurobiological properties, but rather the abstract functional/organizational properties in virtue of which physical state-transitions are systematically appropriate to the content of the mental states they subserve. That is, there is a mathematical level of organization that mediates between intentional mental states and their physical realization. Intentional mental states and state-transitions are realized by certain mathematical states and state-transitions, which in turn are realized by certain physycal states and state-transitions. The mathematical level is the appropriate one for characterizing the abstract system of functional/organizational features that constitutes Nature's engineering design for human cognition.[3]

Thus Marr's tri-level typology for cognitive science is a species of a more generic tri-level typology:

Cognitive State-Transitions. The level of the mental *qua* mental.
Mathematical State-Transitions. The level of functional organization.
Physical Implementation. The level of the physical *qua* physical.

Classicism is one species of this generic three-level approach to cognition. In principle, there could arise one or more alternative, nonclassicist, approaches that are different species of the same genus; more on this below. Examining how classicism itself construes the three levels will be a useful way of articulating the key foundational assumptions of classicism, and of pointing toward alternative possibilities involving different, incompatible, foundational assumptions.

mind for the top level is the specification of and assessment or justification of the function to be computed. How it is computed is to be determined at the middle level – the level of the algorithm.

[3] Some people might doubt whether mathematical state-types are literally instantiated by humans, or by artifacts like computers or neural networks. Those who have scruples about this can, if they like, say that cognitive systems instantiate abstract causal-functional properties – properties that are *indexed* by mathematical state-types, and whose causal-functional relations to one another are mirrored by mathematical relations among the corresponding mathematical state-types.

Classicism supposes that the appropriate mathematical framework for describing cognizers is the theory of computation: mentality is subserved by computational processes, which themselves are physically implemented in the 'wetware' of the brain. On this view of mind, cognitive states are realized mathematically by discrete formal structures, many or all of which have language-like *syntactic* structure. Moreover, the mathematical state-transitions are manipulations and transformations of these representational structures in accordance with programmable rules that are purely formal, i.e., that advert solely to the form or structure of the representations and not to their content. The formal rules governing the representations constitute an *algorithm*: under the relevant assignment of content to the representations, processing in conformity to the rules is guaranteed to effect the cognitive transitions described at the level of cognitive-state transitions.

The contention that cognition is mathematically realized by an algorithm over symbolic representations, i.e., by rule-governed symbol manipulation, involves three basic assumptions:

(1) Intelligent cognition employs structurally complex mental representations.
(2) Cognitive processing is sensitive to the structure of these representations (and thereby is sensitive to their content).
(3) Cognitive processing conforms to precise, exceptionless rules, statable over the representations themselves and articulable in the format of a computer program.

Although (2) implies (1), since you cannot have structure-sensitive processing without structure, these two claims are worth distinguishing explicitly in order to highlight representational structure as well as structure-sensitive processing. Claims (2) and (3) are frequently not distinguished at all or are taken to be equivalent. But although (3) implies (2), (2) does not imply (3) – a fact that will figure importantly below.

Classicists can, and sometimes do, allow that *some* computational processes subserving human cognition might operate on representations with some kind of formal structure other than language-like syntactic structure (e.g., imagistic structure). But they also maintain that the systematic semantic coherence of human thought rests largely, and essentially, on the computational manipulation of representations that encode propositional content via language-like *syntactic* structure. So classicism makes an additional foundational assumption that is more specific than (1):

(4) Many mental representations have syntactic structure.

Assumptions (1)–(4) pertain to the mathematical level of description as conceived by classicism – the level of computation. But to maintain, as classicism does, that cognitive transitions are subserved by an algorithm that computes those transitions is to presuppose something about the top level too, the level of mental states *qua* mental. The transition function being computed, after all, is a *cognitive* transition function (henceforth, CTF), over total cognitive states (henceforth, TCS's). Classicism's presupposition about the cognitive level, so basic that its status as an assumption is often not even noticed, is this:

(5) Human cognitive transitions conform to a *tractably computable* cognitive transition function.[4]

This is hardly a truism; on the contrary, it is a very strong assumption indeed. Even granting that human cognitive transitions conform to a function, there is nothing independent of the assumptions of classicism to indicate that the cognitive transition function must be computable, let alone tractably computable. However, this assumption is never, to the best of my knowledge, argued for on independent grounds. It is obviously presupposed by classicism, since cognitive transitions couldn't possibly be subserved by computation unless those transitions were themselves tractably computable.[5]

Tienson and I call the rules mentioned in (3) *programmable, representation levels rules* (for short, PRL rules). It is important that these rules are supposed to be statable in terms of the representations themselves; i.e., they refer solely to those structural aspects of the representations that play representational roles. Although processing in certain systems may also conform to rules statable at one or more lower levels

[4] The commonly heard term 'tractably computable' is vague and perhaps also somewhat context dependent. But it is not so vague that it cannot do useful work. 'Tractable', in such contexts, means something like capable of being done, or humanly do-able. In different contexts this amounts to computable with roughly the order of computational resources of the human brain, computable by physical devices humans are likely to be able to construct, etc.

It is clear that *tractable* computability, and not computability in the mathematical sense, is at issue, since classical cognitive science is concerned with how human cognitive systems work. There are infinitely many computable functions that are not tractably computable. Consider, for example, any googolplex of unrelated pairings. (A googolplex is 1 followed by 10 to the 100th 0's.) The difference between infinite and finite but huge is unimportant for cognitive science.

[5] I am not here assuming that classicism is committed to cognitive transitions that are deterministic. Classicists can, and often do, provide for non-deterministic cognitive transitions, by building 'dice throws' into the underlying algorithms. The notion of a cognitive transition function can be understood so that the CTF need not be deterministic. The operative notion of 'computing a function' needs to be understood accordingly; cf. Horgan and Tienson (1996), Chapter 2.

(e.g., the level of machine language in a conventional computer), such lower-level rules are not the kind that count. (Henceforth, whenever I talk about classicism's contention that cognition is subserved by computation, I will mean computation *over representations*, i.e., processing that conforms to PRL rules.[6]) This is important because there can be (non-classical) systems which conform to rules at lower levels, but which do not conform to rules statable only at the level of representations, as I discuss in Section 9 below.

Classicism does not assert that the PRL rules of cognitive processing must be represented by (or within) the cognitive system itself. Although programs are explicitly represented as stored data structures in the ubiquitous general-purpose computer, stored programs are not an essential feature of the classical point of view; a classical system can conform to representation-level rules simply because it is hardwired to do so. It is, for example, plausible from the classical point of view to regard some innate processes as hardwired.

Claims (1)–(5), I take it, are the foundational assumptions of classicism. In terms of logical relations, (3) is most basic; for, (3) implies (1), (2), and (5), while (4) adds the contention that the relevant structure in the representations is largely, if perhaps not entirely, syntactic. But in another way, (5) is most basic; for the computational approach to the mind can't possibly be right unless human cognitive transitions themselves are tractably computable.

Classicism is a "just-so story" about the mind. By this I mean on the one hand that it attempts to tell a fairly general and comprehensive story about how human cognition might work, and on the other hand that the quality and quantity of evidential support for this conception of human mentality is fairly meager. Moreover, it is hardly the case that the evidence in support of classicism has been mounting steadily and inexorably. On the contrary, insofar as AI-style cognitive science has attempted to progress beyond the modeling of fairly simple, highly circumscribed, cognitive, processes, the results typically have been frustrating and disappointing.

2. Troubles for Classicism[7]

By the early 1980's, certain kinds of difficulties were arising quite

[6] The term 'computation' is vague. Some researchers seem inclined – not necessarily inappropriately – to use the term considerably more broadly than I do here. But the sense specified in the text is surely the sense of 'computation' relevant to discussions of classicism. If, for example, there are physical systems – connectionist networks, or human brains, perhaps – that effect transitions that do not correspond to a computable function (in the sense of the Church-Turing thesis), what they do is not computation in the classical sense.

[7] This section is drawn largely from section 2 of Horgan and Tienson (1994).

persistently and systematically within classicism. Examination of these difficulties makes it seem quite likely that they are difficulties in principle, stemming from the fundamental assumptions of the framework, and not mere temporary setbacks.

The difficulties center largely on what is called the *frame problem*. In *The Modularity of Mind*, Jerry Fodor (1983) characterizes the frame problem as "the problem of putting a 'frame' around the set of beliefs that may need to be revised in light of specified newly available information" (112–13). In the closing pages of that work, Fodor argued that the problems look to be in-principle ones, and hence that the prospects for understanding human cognitive processes like belief fixation within the framework of classical cognitive science are very bleak indeed.

The main claim of *The Modularity of Mind* is that the human cognitive system processes a number of important subsystems that are *modular*: domain specific, mandatory, limited in their access to other parts of the larger cognitive system, fast, and informationally encapsulated. There is good evidence, Fodor argues, that human input systems, including those that mediate visual perception, exhibit modularity. Where classicism has gotten somewhere, he says, it is in understanding such modular subsystems, which by their nature delimit the class of relevant information.

Classicism has made very little progress in understanding *central* processes, however. Belief fixation – the generation of new beliefs on the basis of current input together with other beliefs both occurrent and non-occurrent – is a paradigmatic example. Updating of the overall belief system, in light of currently available new information, is another closely related example. Fodor argues convincingly that these processes are non-modular: they need to have access to a wide range of cognitive subsystems, and to information on an indefinitely wide range of topics. And the very considerations that point to non-modularity, he maintains, also constitute grounds for extreme pessimism about the prospects for explaining central processes within the framework of classical computational cognitive science.

Fodor articulates these considerations in terms of the analogy between belief fixation in human cognition and scientific confirmation. He writes:

> Central systems look at what the input systems deliver, and they look at what is in memory, and they use this information to constrain the computation of 'best hypotheses' about what the world is like. The processes are, of course, largely unconscious, and very little is known about their operation. However, it seems reasonable enough that something can be inferred about them from what we know about *explicit* processes of nondemonstrative inference – viz., what we know about empirical inference in science. (104)

Scientific confirmation, "the nondemonstrative fixation of belief in science," has two crucial features. It is (in Fodor's terminology) *isotropic* and *Quineian.* He says:

> By saying that confirmation is isotropic, I mean that the facts relevant to the confirmation of a scientific hypothesis may be drawn from anywhere in the field of previously established empirical (or, of course, demonstrative) truths. Crudely: everything that the scientist knows is, in principle, relevant to determining what else he ought to believe . . .
>
> By saying that scientific confirmation is Quineian, I mean that the degree of confirmation assigned to any given hypothesis is sensitive to properties of the entire belief system; as it were, the shape of our whole science bears on the epistemic status of each scientific hypothesis. (105–7)

Being isotropic and being Quineian are two ways in which confirmation is holistic. Any bit of actual or potential information from any portion of the belief system might, in some circumstances, be evidentially relevant to any other. And confirmation is relative to the *structure* of the whole of the current belief system and of potential successor systems.

Since belief fixation in human cognition is a matter of inductive inference from the information provided by input systems and the information in memory, evidently it too must be isotropic and Quineian. Fodor concludes that it must be non-modular. He also stresses that these global aspects of belief fixation look to be at the very heart of the problems that classicism has encountered in attempting to understand such central processes:

> The difficulties we encounter when we try to construct theories of central processes are just the sort we would expect to encounter if such processes are, in essential respects, Quineian/isotropic . . . The crux in the construction of such theories is that there seems to be no way to delimit the sorts of informational resources which may affect, or be affected by, central processes of problem-solving. We can't, that is to say, plausibly view the fixation of belief as effected by computations over bounded, local information structures. A graphic example of this sort of difficulty arises in AI, where it has come to be known as the "frame problem" (i.e., the problem of putting a "frame" around the set of beliefs that may need to be revised in light of specified newly available information). (112–3)

One classical research strategy has been to artificially delimit necessary informational resources by considering artificially simple 'worlds' or artificially circumscribed task domains. Such approaches have notoriously failed to 'scale up' to more complex cases in more realistic task domains.

And when one considers the sorry history of attempts in philosophy of science to construct a theory of confirmation, the prospects for understanding central processing within the classical computational paradigm look very discouraging indeed:

> Consider . . . the situation in the philosophy of science, where we can see the issues about fixation of belief writ large. Here an interesting contrast is between deductive logic – the history of which is, surely, one of the great success stories of human history – and confirmation theory which, by fairly general consensus, is a field that mostly does not exist. My point is that this asymmetry, too, is likely no accident. Deductive logic is the theory of validity, and validity is a *local* property of sentences. Roughly, the idea is that the validity of a sentence is determined given a specification of its logical form, and the logical form of a sentence is determined given a specification of its vocabulary and syntax. In this respect, the level of validity contrasts starkly with its level of confirmation, since the latter . . . is highly sensitive to global properties of belief systems . . . The problem in both cases is to get the structure of the entire belief system to bear on individual occasions of belief fixation. We have, to put it bluntly, no computational formalisms that show us how to do this, and we have no idea how such formalisms might be developed . . . In this respect, cognitive science hasn't even *started;* we are literally no farther advanced than we were in the darkest days of behaviorism . . . If someone – a Dreyfus, for example – were to ask us why we should even suppose that the digital computer is a plausible mechanism for the simulation of global cognitive processes, the answering silence would be deafening. (128–9)

These are wise words. Let me underscore their wisdom by dwelling just a bit on the depth of, and the apparently in-principle nature of, the difficulties encountered by attempts to model global cognitive processes computationally. Take, first, the Quineian aspect of belief systems. Simplicity and conservatism are properties of (or relations between) belief systems that depend upon the formal, semantic, and evidential relations among *all* of the beliefs in the system(s). A computational system would have to somehow survey the entire stock of beliefs, in a manner that tracks all the multifarious interconnections among the beliefs, and somehow derive a *measure* of net overall simplicity and net overall conservatism from these local features. As Fodor said, "We have . . . no computational formalisms that show us how to do this, and . . . no idea how such formalisms might be developed . . ."

In addition, when new information comes in from the input modules, the central system would have to find, from among the vastly many competing, incompatible ways of revising the whole system to accommodate this new information, a mode of revision that maintains overall simplicity and conservatism better than most of the others. All this would have to be done via *tractable* computation, executable quite rapidly. Not only do we have no computational formalisms that show us how to do this; it's a highly credible hypothesis that a (tractable) computational system with these features is just impossible, for belief systems on the scale possessed by human beings.

Now consider isotropy – the potential relevance of anything to anything. Take, for example, the problem of fetching items from

memory that are relevant to determining whether a new putative belief should be accepted or rejected. For a belief system of any size, of course, it is utterly inefficient to search through each bit of old information to see whether and how it is relevant. The problem here is sometimes thought of as a problem of representing common-sense information – i.e., representing it in a way that leads to access when relevant – and sometimes as the problem of content addressable memory. But these labels really obscure the extent of the difficulty. Does the system first identify a bit of information stored in memory as relevant, and then retrieve it? But how can it identify the information as relevant until it is found? Apparently the system must find the information in memory and identify it as relevant all at once. As Hume said, "One would think that the whole intellectual world of ideas was at once subjected to our view, and that we did nothing but pick out such as were most proper for our purposes" (*Treatise*, I, I, vii, 24). What seems to be needed is *relevance* addressable memory, whatever that might be. Once again, we lack the slightest clue how a tractable computational process could accomplish this for belief systems of the size possessed by humans. Indeed, it seems entirely likely that it can't be done via (classical) computation at all. From the point of view of computationalism, Hume's succeeding remark is quite apt: those ideas "are thus collected by a kind of magical faculty in the soul, which . . . is however inexplicable by the utmost efforts of human understanding."[8]

The moral of such considerations, as Fodor comes right to the brink of saying in the passage most recently quoted above, is that human central processing is apparently too subtle and complex to be subservable by representation-level computation – i.e., by PRL rules.

3. Connectionism: A New Game in Town

Connectionism emerged as a large-scale research program in the 1980's, largely in response to the recurrent, recalcitrant, difficulties that led Fodor to his bleak conclusions about classicism's prospects for modeling central cognitive processes.

A connectionist system, or neural network, is a structure of simple neuron-like processors called nodes or units. Each node has directed connections to other nodes, so that the nodes send and receive excitatory and inhibitory signals to and from one another. The total input to a node determines its state of activation. When a node is on, it sends out signals to the nodes to which it has output connections, with the inten-

[8] Tienson and I elaborate frame-type problems more fully in Horgan and Tienson (1994, 1996). We emphasize that the difficulties actually go even deeper, because relevance is not only isotropic but also Quineian. We offer additional, related, arguments against cognition as conforming to PRL rules in Horgan and Tienson (1988, 1989, 1996).

sity of a signal depending upon both (i) the activation level of the sending node and (ii) the strength or "weight" of the connection between it and the receiving node. Typically at each moment during processing, many nodes are simultaneously sending signals to others.

When neural networks are employed for information processing, certain nodes are designated "input" units and "output" units, and potential patterns of activation across them are assigned interpretations. (The remaining nodes are called "hidden units.") Typically a "problem" is posed to a network by activating a pattern in the input nodes; then the various nodes in the system simultaneously send and receive signals repeatedly until the system settles into a stable configuration; the semantic interpretation of the resulting pattern in the output nodes is what the system currently represents, hence its "answer" to the problem.

Connectionist systems are capable of "learning" from "experience" by having their weights changed systematically in a way that depends upon how well the network has performed to date in generating solutions to problems posed to it as a training regimen. (Typically the device employed is not an actual neural network, but a simulation of one on a standard digital computer.) Investigating various weight-change algorithms has been a prominent part of connectionist research. Networks are "trained up" by adjusting weights in ways that depend upon, and gradually improve, the system's performance in a given task domain.

The most striking difference between such networks and conventional computers is the lack of an executive component. In a conventional computer the behavior of the whole system is controlled at the central processing unit (CPU) by a stored program. A connectionist system lacks both a CPU and a stored program. Nevertheless, often in a connectionist system certain activation patterns over sets of hidden units can be interpreted as internal representations with interesting content, and often the system also can be interpreted as embodying, in its weighted connections, information that gets automatically accommodated during processing without getting explicitly represented via activation patterns.

Connectionist models in cognitive science have yielded particularly encouraging results for cognitive processes like learning, pattern recognition, and so-called multiple-soft-constraint satisfaction (i.e., solving a problem governed by several constraints, where an optimal solution may require violating some constraints in order to satisfy others – e.g., successfully classifying a given three-legged animal as a dog, even though dogs have four legs). For instance, Terry Sejnowksi and Charles Rosenberg (1987) trained a network they called NETtalk to convert inputs that represent a sequence of letters, spaces, and punctuation constituting written English, into outputs that represent the audible sounds constituting the corresponding spoken English. (The phonetic

output code then can be fed into a speech synthesizer, a device that actually produces the sounds.)

In a connectionist system, information is actively represented as a pattern of activation. When the information is not in use, that pattern is nowhere present in the system; it is not stored as a data structure. The only representations ever present are the active ones. On the other hand, information can be said to be *tacitly* present in a connectionist system – or "in the weights," as connectionists like to say – if the weighted connections subserve *representation-level dispositions* that are appropriate to that information. As Tienson and I have lately begun to put it (Horgan and Tienson 1995, 1996), such information constitutes *morphological* content in the system (i.e., content that is present in virtue of the system's structure), rather than explicitly-represented content. Some, but not necessarily all, of this morphological content is also present in the system in another way as well: it is *potentially* active. I.e., the weighted connections dispose the system to generate, in appropriate circumstances, active representations with this information as their content. Among the apparent advantages of connectionist systems, by contrast with classical computational systems, is that morphological information "in the weights" gets accommodated automatically during processing, without any need for a central processing unit to find and fetch task-relevant information from some separate memory banks where it gets stored in explicit form while not in use.

Remembering, on this view, is not a matter of retrieving a representation from a mental file cabinet called memory. Rather, memory information is a species of morphological content. Memories are not fetched from storage, but rather are literally *recreated* over and over again by whatever internal or external stimuli remind you of them. Learning too is conceived quite differently within connectionism than it is within classicism, since connectionist systems do not store representations. Because learning involves the system's undergoing weight changes that render its representation-forming dispositions appropriate to the content of what is learned, learning is the acquisition, "in the weights," of new morphological content.

Connectionist modeling is now a major branch of cognitive science, along side traditional AI-style cognitive modeling. (The two approaches also are sometimes combined, in so-called "hybrid" systems.) But these are early days yet for connectionism; most extant models address comparatively simple cognitive tasks, and it remains to be seen how well the connectionist approach will "scale up" when it comes to designing models with capacities that more closely approximate the cognitive capacities of humans. More specifically, it remains to be seen whether connectionism will be able to make inroads against the recalcitrant, in-principle looking, problems that plague classicism, like the frame problem.

4. Non-Sentential Computationalism: Another "Just-So Story"

How exactly might a connectionism-inspired approach to mentality differ from classicism, at the level of foundational assumptions about cognition? At present, connectionism is primarily an alternative way of *doing things* in cognitive science, rather than an alternative set of doctrines or theses: one constructs network models and trains them up with learning algorithms, rather than writing programs. There is really no such thing as "the connectionist conception of the mind," in the form of a determinate set of foundational assumptions differing in specific, explicit, ways from those of classicism. Philosophers have discussed several, incompatible conceptions of mind that could develop from or be wedded to connectionism.

The least radical possibility for connectionism is that it might provide for new ways of implementing computational processes over structurally complex mental representations, including language-like representations. This might lead to interesting new developments at higher levels of description – for instance, to new kinds of algorithms, ones that are more naturally implemented in connectionist architecture than in conventional computers. Important as such new developments might be, however, they would not amount to a new conception of cognition, over against classicism's conception. On the contrary, as long as assumptions (1)–(5) from section 2 remain in force, cognition is still being conceived classically. Implementational connectionism would be no deviation from classicism at all, but merely a species of it.[9]

One way to depart from classicism, however, is to reject assumption (4) of the five foundational assumptions of classicism I cited in Section 1 – viz., the assumption that many mental representations have syntactic structure – while still retaining the others. On this view of the mind, which Tienson and I call *non-sentential computationalism* (henceforth, NSC), human cognitive transitions do conform to a tractably computable transition function over total cognitive states; and these transitions are still subserved by an algorithm that computes that function – i.e., by PRL rules over mental representations. However, the representations themselves are not sentential, and thus their structure is not *syntactic* structure. NSC rejects the "language of thought" to which classicism is committed, and a number of philosophers are attracted to connectionism because they see it as supporting a denial of the language of thought.

With connectionist networks in mind one might claim, for instance, that mental representations are non-sentential *activation vectors* and that the algorithm implemented by a network effects vector-to-vector transformations that are systematically sensitive to vectorial structure,

[9] Cf. Fodor and Pylyshyn (1988). For a more detailed discussion of the matter of "mere implementation," see Horgan and Tienson (1992).

and thereby are systematically sensitive to the content of these non-sentential representations. (A vector is essentially just an ordered n-tuple of items; an activation vector is an ordered n-tuple of activation values of specific nodes of a neural network.) Connectionist networks, on this view of things, are devices that are naturally suited to perform such non-sentential computation; vectors are realized as activation patterns over sets of nodes. Thus, the claim goes, connectionism can and should evolve toward non-sentential computationalism as an alternative to classicism. A number of philosophers appear to think that connectionism, insofar as it purports to provide an alternative to classicism rather than merely a new implementation of it, amounts to NSC. This includes philosophical fans of connectionism like Paul Churchland (1989, 1995), and philosophical foes like Fodor and Pylyshyn (1988).

NSC is thus a commonly held foundational interpretation of connectionism, an interpretation that construes connectionist models as employing a non-classical kind of representational-level computation that (i) is highly parallel (rather than serial, as in conventional computers), (ii) eschews the distinction between a central processor and separate memory banks for information storage, and (iii) eschews language-like representations with syntactic structure. Feature (iii) is the most fundamental aspect of this non-classical foundational interpretation, since classicism *per se* is not officially committed either to serial cognitive architectures as opposed to parallel ones, or even to architectures in which representations remain explicitly present in separate memory-repositories when they are not figuring in active processing.

NSC is not obviously a correct foundational interpretation of all extant connectionist models, however. On the contrary, there are certain models for which NSC is *prima facie* not correct – such as those of Berg (1992), Pollack (1990), Legendre, Miyata, and Smolensky (1991), and Smolensky (1990, in press) that are naturally interpreted as involving representations with syntactic structure. Nor is NSC obviously the only possible, or the most natural, or the most attractive, foundational framework for connectionist cognitive science. There is another kind of departure from assumptions (1)–(5) of classicism that potentially could mesh well with connectionism, and that deserves serious consideration and exploration. I will turn to that in the later parts of the paper.

NSC, like classicism, is itself a "just-so story" about the mind; for NSC too, the quality and quantity of evidential support is fairly meager. For one thing, since connectionism is still in its early days, it remains to be seen whether or not connectionist approaches will fare any better than classicism when it comes time to scale up from extremely simple models to the modeling of more complex, more thoroughly human-like,

cognitive processes. It is possible that only models like those mentioned in the preceding paragraph will scale up. But furthermore, there are persuasive reasons for thinking that if the connectionist approach is to manifest any serious promise for explaining aspects of human mentality that have resisted explanation within classicism, this approach will need to be wedded to a foundational conception of mentality different from *either* classicism or NSC. I will turn to these reasons shortly, after briefly discussing the impact of NSC on the debate about "folk psychology" within philosophy of mind.

5. Connectionism and Folk Psychology

Some philosophers have recently argued that if human mentality were to be explainable by means of the connectionist approach, then this would show that common-sense belief/desire psychology – so-called "folk psychology" – is radically false. I will briefly consider two influential arguments of this kind, each of which appears to presuppose that NSC is the right foundational interpretation of connectionism.

Paul Churchland (1988) argues essentially this way: Since (1) connectionism allegedly eschews language-like representations, and (2) folk psychology allegedly is committed to a "sentential kinematics," connectionism is incompatible with folk psychology.

One reply to this line of argument, which I myself favor (Horgan and Graham 1991, Horgan 1993), goes as follows: Folk psychology per se is not committed to language-like mental representations; rather, the language-of-thought (LOT) hypothesis is one particular empirical hypothesis about how beliefs and desires are *realized* in humans, at an abstract cognitive-scientific level of description. If the LOT hypothesis should turn out to be false, then this would show not that humans lack beliefs and desires, but rather that beliefs and desires are realized some *other* way at the functional-mathematical level of description.[10]

William Ramsey, Stephen Stich, and Joseph Garon (1990) offer a different argument for the incompatibility of connectionism and folk psychology, roughly this: (1) Folk psychology is committed to the possibility that a person can have two different reasons for an action (two different belief/desire combinations that rationalize it), one of which causally generates the action and the other of which does not. (Ramsey, Stich, and Garon call this "modularity," using the term somewhat differently than does Fodor.) (2) If human cognition works in the manner of recent connectionist models, then this commitment of folk

[10] The claim that connectionism eschews sentential representations is tendentious. Below I describe a foundational construal of connectionism, different from NSC, that does not necessarily repudiate syntax. Indeed, in the version of that construal that Tienson and I favor, syntax gets retained – although it is conceived quite differently than in classicism.

psychology is false. Hence (3) if connectionism is right, then people have no beliefs and desires.[11]

One reply to this argument, which I myself favor (Horgan and Tienson 1995), may be briefly summarized as follows. The argument fails in three different ways. First, folk psychology is clearly committed only to cases of modularity where the causally active reason is *occurrent*, whereas the causally dormant one remains merely *dispositional*; and connectionism accommodates this possibility just fine. Second, even if folk psychology were committed to other forms of "modularity" that turned out to be precluded under connectionism, this kind of false commitment would not mean that folk psychology is so radically false that there are not really beliefs and desires; rather, it would only mean that folk psychology is somewhat mistaken about the nature of beliefs and desires. Third, connectionism actually can accommodate various other forms of "modularity" anyway; these forms would be manifested at the abstract functional-mathematical level of description, rather than in *physically* discrete states (see Section 2.5 of Horgan and Tienson 1995).

6. Troubles for Non-Sentential Computationalism

Fodor and Pylyshyn (1988) set forth an influential critique of connectionism. In effect they construe the connectionist approach to mentality, insofar as it purports to be an alternative to classicism rather than merely a new "implementation architecture," as NSC. They argue that connectionism is a serious step backward, a form of recidivism in cognitive science. The only known way to approach the task of modeling the (large-scale) semantic coherence of thought, they claim, is to invoke processing that employs language-like representations that encode propositional content syntactically and are subjected to processing which, by virtue of being suitably structure-sensitive, is *thereby* suitably content-sensitive. By giving up on syntactic structure, they argue, connectionism in effect embraces *associationism* – by which they mean a view of mental processing in which the system's processing is primarily driven by its sensitivity to statistical regularities among the items its internal states represent. And the problems repeatedly faced by associationism in the history of psychology, they claim, have amply shown it to be seriously inadequate – mainly because of its limited capacity to explain complex, semantically coherent, thought processes. They regard the apparent link with associationism as grounds for maintaining that connectionism, insofar as it strives to do

[11] Davies (1991) has argued similarly. Davies' argument appeals to certain alleged commitments of folk psychology involving folk-psychological concepts other than the concept of belief.

more than implement classicism, is bound to founder because of the same kinds of problems that plagued traditional associationism. Concerning the non-implementational prospects for connectionist networks in cognitive science, they write:

> A good bet is that networks sustain such processes as can be analyzed as the drawing of statistical inferences; as far as we can tell, what network models really are is just analog machines for computing such inferences. Since we doubt that much of cognitive processing does consist of analyzing statistical relations, this would be a quite a modest estimate of the prospects for network theory. (1988, 68)

Closely related to Fodor and Pylyshyn's charge that connectionism (construed as NSC) constitutes an unpromising step backwards in cognitive science is the following, more general worry about NSC. What exactly are we supposed to be gaining, in terms of our abilities to model cognitive processes, by adopting an approach which (i) retains the assumption that cognitive processing is representation-level computation, but (ii) eschews one extremely powerful way to introduce semantic coherence into representation-level computation: viz., via the syntactic encoding of propositional content? Qua representation-level computation, it looks as though this amounts to trying to model semantically coherent thought processes with one hand – the *good* hand – tied behind one's back. What seems needed, in order to successfully address recalcitrant problems that have plagued classicism (like the frame problem), is a non-classical approach that is more powerful than classicism. Yet NSC, by retaining the view that cognition is subserved by representation-level computation and then repudiating syntactic structure in the representations, evidently offers us something *less* powerful, viz., a seriously crippled cousin of classicism.

A fan of NSC might be expected to reply this way to the claim that NSC is just crippled classicism:

> But connectionist models get by without any explicit stored memories, with lots of information 'in the weights'; and networks aren't programmed the way classical systems are.

But here is the appropriate counter-reply:

> To the extent that the processing conforms to PRL rules, it seems we could get these same features in a classical system in which (i) all the rules are hardwired rather than explicitly represented, and (ii) lots of information is implicitly accommodated in the (hardwired) rules rather than being explicitly stored in memory. We could get parallelism too, e.g., in a parallel production-system architecture in which all applicable production-rules fire simultaneously.

This counter-reply underscores the fundamental problem: NSC is evidently just a limited variant of classicism. It is a variant because it continues to hold that cognition is subserved by processes that conform to PRL rules; and it is limited because it restricts itself to forms of computation that do not employ either syntactically structured representations or explicitly stored representations.

Classicism began to encounter recalcitrant problems when it attempted to move beyond artificially simple, informationally encapsulated, cognitive tasks; classicist models persistently failed to "scale up" in a way that would explain complex and holistic cognitive processes like human belief-fixation. Given that NSC is evidently a less powerful cousin of classicism, the threat of similar scale-up problems looms very large for NSC as well.

7. Taking Stock: Seeking Out Another "Just-So Story"

Given how serious are the apparent problems facing both classicism and NSC, there is ample reason to seek out a foundational approach to cognitive science that would differ substantially from both. Those problems, as canvassed in sections 2 and 6 above, suggest a likely source of the limitations of both classicism and NSC: the assumption that cognition is subserved by representation-level computation, and the assumption that cognitive transitions conform to a tractably computable transition-function. These are assumptions (3) and (5) respectively, from the list in section 1 of classicism's five key foundational assumptions.

The problems threatening both classicism and NSC do not, however, pose any immediate or obvious threat to the other three foundational assumptions of classicism: structurally complex mental representations, structure-sensitive processing, and syntactic structure. Nor do those difficulties call into question the generic tri-level typology for cognitive science I mentioned in section 1: the cognitive level of description, the functional-mathematical level, and the physical level, with states at each level realized by lower-level states.[12]

[12] Thus, they do not call into question the contention that cognition employs *representations*; it is the "rules" part of the so-called "rules and representations" paradigm that is the source of the problem.

There is, however, an interpretation of connectionism claiming that connectionist models do not really employ internal representations at all in their hidden units (and *a fortiori*, do not employ internal representations with language-like structure). This view has been defended – for example, by Rodney Brooks (1991) – on the grounds that putative representations in connectionist systems play no genuine explanatory role. It has also been defended – for instance, by Hubert and Stuart Dreyfus (1990) – on the basis of a Heideggerian critique of the notion of mental representation itself. This approach goes contrary to the views of most (but not all) practising connectionists, who typically

So a natural path to pursue, in seeking out an alternative foundational approach to cognitive science, would be along the following lines. On one hand, the approach would retain some key features of other approaches. First, it would retain the generic tri-level typology, with the associated idea that Nature's cognitive engineering involves a system of physically realizable mathematical states, plus a suitably systematic cognitive/mathematical realization relation, such that mathematical state-transitions systematically subserve content-appropriate cognitive state-transitions. Second, it would retain the idea that there is rich structure in the system of mathematical states, structure that figures both in mathematical state-transitions (structure-sensitive processing) and in the systematicity of the realization relation. Third, it would retain, or at least would be receptive to, the idea that such functional-mathematical structure includes some form of syntax, so that the system of representations qualifies as a language of thought. In all these respects, it would resemble classicism.

On the other hand, at the mathematical level of description the alternative approach would invoke some form of mathematics more powerful than the discrete mathematics of computation theory – some mathematical framework that yields richer kinds of functional-mathematical structure. Correlatively, the approach would invoke a cognitive/mathematical realization relation that systematically links cognitive and mathematical states in a way that suitably harnesses the greater richness of mathematical structure. The hoped-for result would be that mathematical state-transitions systematically subserve cognitive state-transitions that are too rich and too subtle to conform to any tractably computable cognitive-level transition function.

In fact there is a body of mathematical concepts, techniques, and results that has some promise for filling this tall order, and that goes naturally with connectionism: dynamical systems theory. I turn next to that.

8. Connectionist Networks and the Mathematics of Dynamical Systems

If one focuses on the theoretical and mathematical aspects of connectionist modeling in cognitive science, one finds connectionists increasingly invoking the mathematics of dynamical systems. It is also becoming increasingly important in cognitive neuroscience; cf. Churchland (1988), Amit (1989), Skarda and Freeman (1987), and Freeman (1991). The mathematics of dynamical systems is fundamentally continuous mathematics rather than discrete mathematics – even though it can be brought to bear on systems which, at the relevant level of description, are literally discrete (say, because they evolve in discrete time-steps).

To treat a system as a dynamical system is to specify in a certain way

its temporal evolution, both actual and hypothetical. The set of all possible states of the system – so characterized – is the system's abstract, high-dimensional *state space*. Each magnitude or parameter of the system is a separate dimension of this space, and each possible state of the system, as determined by the values of these parameters, is a point in its state space.

In classical mechanics, for instance, the magnitudes determining the state space of a given mechanical system are the instantaneous positions, masses, and velocities of the various bodies in the system. Temporal evolution of such a system, from one point in its state space to another, is determined by Newton's laws, which apply globally to the entire state of the system at an initial time.[13]

Connectionist systems are naturally describable, mathematically, as dynamical systems. The magnitudes determining the state space of a given connectionist system are the instantaneous activation levels of each of the nodes in the network. Thus the state space of a network is frequently called its "activation space". The activation space of a network has as many dimensions as the network has nodes. The rule is governing the system's temporal evolution apply locally, at each node of the system; this simultaneous local updating of all nodes determines the system's evolution through time.

A dynamical system, as such, is essentially the full collection of temporal trajectories the system would follow through state space – with a trajectory emanating from each possible point in state space. A dynamical system can be identified with a set of state space trajectories in roughly the same sense in which a formal system can be identified with the set of its theorems. Just as there are many different axiomatizations of a formal system such as the modal logic system S5 or the classical propositional calculus, so likewise there might be many different mathematical ways of delineating a certain dynamical system or class of dynamical systems. (It is common practice to use the term 'dynamic system' ambiguously for abstract mathematical systems and for physical systems – such as planetary systems and certain networks – whose behavior can be specified via some associated mathematical

posit internal representations in connectionist models and assign them a central explanatory role. For a critique of anti-representationalist construals of connectionism, see Clark and Toribo (1994).

[13] Some dynamical systems, including many that have been studied in physics, evolve in accordance with relatively simple laws, typically involving global states of the system. Usually such laws are expressed via differential equations, or difference equations if discrete time-steps are involved. But it is very important to appreciate that in general, a dynamical system need not conform to such laws, and in particular, connectionist networks need not conform to simple update-rules expressible over *global* activation-states.

dynamical system. Here I use the term mainly for mathematical systems.)

A dynamical system can be thought of as a geometrical/topological mathematical object. A useful geometrical metaphor for dynamical systems is the notion of a *landscape*. Consider a system involving just two magnitudes; its associated state space is two dimensional, and thus can be envisioned as a Cartesian plane. Now imagine this plane being topologically molded into a contoured, non-Euclidean, two dimensional surface. Imagine this "landscape" oriented horizontally, in three dimensional space, in such a way that for each point *p* in the system's two dimensional state space, the path along the landscape that a ball would follow if positioned at *p* and then allowed to roll freely is the temporal trajectory that the system itself would follow, through its state space, if it were to evolve (without perturbation) from *p*. A dynamical system involving n distinct magnitudes can be thought of as a landscape too; it is the *n*-dimensional analog of such a two dimensional, non-Euclidean, contoured surface: i.e., a topological molding of the n-dimensional state space such that, were this surface oriented 'horizontally' in an (*n*+1) dimensional space, a ball would "roll along the landscape," from any initial point, in a way that corresponds to the way the system itself would evolve through its state space (barring perturbation) from that point.

In connectionist models, cognitive processing is typically construed as the system's evolution along its activation landscape from one point in activation space to another – where at least the beginning and end points are interpreted as realizing intentional states. When a problem is posed to the system by activating a set of nodes which are interpreted as having a certain content, from a dynamical systems perspective this amounts to positioning the system at some point in its state or activation space. (Which specific point this is will often depend also on the current activation level of nodes other than those involved in posing the problem.) The network eventually settles into a stable state which constitutes its 'solution' to the problem – another point in its activation space.

Connectionist networks 'learn' by progressive incremental changes in the weights, bringing about appropriate changes in the system's trajectories through activation space. Hence it is natural to think of learning in a connectionist network as *molding* its activation landscape.[14]

[14] Learning is also typically construed as temporal evolution of a connectionist network through a state space. When the issue is learning as opposed to processing, however, the weights on the network's connections are viewed as malleable rather than fixed; learning involves changes in the weights, constituting a trajectory through *weight* space.

9. Dynamical Cognition: A Third "Just-So Story"

John Tienson and I have proposed a non-classical foundational framework for cognitive science that differs in important ways from both classicism and NSC (Horgan and Tienson 1994, 1996). This alternative approach is inspired partly by the emergence of the connectionist movement, partly by the above-mentioned problems facing both classicism and NSC, and partly by the natural links between neural networks and dynamical systems theory. We call it the *dynamical cognition* framework (for short, the DC framework) because of the central role it assigns to the mathematics of dynamical systems.

The DC framework is a non-classical instantiation of the generic trilevel typology for cognitive science I described in Section 1, in which the mental and physical levels are mediated by a functional-mathematical level of description, a level that is thus the locus of Nature's cognitive design. Two ideas are central. First, at the mathematical level a cognitive system is a high-dimensional dynamical system, subservable by a neural network. Each mathematical state in the dynamical system – i.e., each point in its state space – corresponds to a distinct *total activation state* of the neural network. The dynamical system as a whole is thus a high-dimensional activation landscape, and mathematical state-transitions are temporal trajectories along this landscape.

Second, the cognitive/mathematical realization relation pairs *total cognitive states* (for short, TCS's) with points on the activation landscape. Thus the points that realize TCS's are the cognitive system's *representations*.[15] Since a point on the activation landscape corresponds to a total activation state of the neural network, TCS's and the points that realize them are physically realized in a fully "distributed" way: by total physical states of the network.[16]

As I explained in Section 7, the apparent moral of the problems with classicism and with NSC is that human cognitive transitions are evidently too subtle and too complex to conform to any tractably computable transition function over cognitive states. What is wanted, then, is a foundational approach more powerful than approaches that invoke representation-level computation: an approach that provides richer kinds of mathematical structure, plus a cognitive/mathematical

[15] As in classicism and in NSC, the notion of a representation is effectively a multilevel notion. A representation is a physically realized functional-mathematical state which itself realizes some intentional cognitive state.

[16] In many connectionist models, activation vectors over some but not all of the nodes in the network are treated as representations. From the perspective of the DC framework, what this amounts to is that a given intentional state is multiply realizable mathematically by various different points in activation space, viz., each point that has the spatial coordinates specified in the vector.

realization relation under which this mathematical structure gets suitably exploited in Nature's cognitive engineering. With this goal in mind, let us consider in turn each of the two core features of the DC framework: the role it assigns to the mathematics of dynamical systems, and its treatment of the cognitive/mathematical realization relation.

In the DC framework, mathematical structure in the system of representations is quite different than in classicism. The mathematical states (viz., points in a high-dimensional space) lack intrinsic mathematical structure; in this respect they differ from the mathematical states that serve as representations in classicism. Hence the intrinsic physical aspects of the physical state (viz., a total activation state of a neural network) that realizes a mathematical state do not realize intrinsic mathematical structure – since a point doesn't *have* any intrinsic mathematical structure.[17]

Nevertheless, there is extremely rich mathematical structure in the system of representations. It is a non-intrinsic, *relational* structure among the representational points, involving various ways that these points are systematically positioned relative to each other on the high-dimensional activation landscape. The range of positional relations potentially available to serve this role is enormous. For instance, there are numerous kinds of distance relations between points: not only overall distance in n-dimensional activation space, but also the vastly many lower-dimensional distance relations in the various sub-spaces of activation space. In addition, there are innumerable relative-position relations involving the topography of the landscape itself: for example, two points on the landscape might be position-similar by both being on a common incline on the landscape. Such topography-based relations can be very intricate, because the activation landscape of a neural network can exhibit enormously complex, enormously non-homogeneous, variation in local topography.[18]

[17] A vector, of course, does have mathematical structure: it is an ordered n-tuple. And as I remarked in Section 4, NSC tends to regard the intrinsic structure of vectors as the principal sort of mathematical structure that figures in Nature's cognitive engineering. But from the perspective of the DC framework, an activation vector is merely a specification of the spatial coordinates of a point (or a class of points) in activation space.

[18] This topological complexity can involve, among other factors (here I employ the lingo of dynamical systems theory), the nature of the "attractors," including "chaotic" ones; the nature of the "basin boundaries," including "fractal" ones; and highly intricate intertwinings of "attractor basins," some of them fractal. Topological complexity tends to be especially prevalent in nonlinear dynamical systems; and the dynamics of standard neural networks is *nonlinear*, since the nodes typically update their activation in accordance with some nonlinear function. There are numerous recent books discussing dynamical systems, intended for the general audience and highlighting dramatic recent developments involving chaos, fractals, and nonlinearity. See, for instance, Gleick (1987), Stewart (1989).

At the physical level of description, the rich non-intrinisic mathematical structure in the representational system is embodied in a thoroughly non-intrinsic way. It is not embodied in single activation states of the neural network (since each total activation state corresponds mathematically to a structureless point). Nor is it embodied in the standing, intrinsic, physical structure of the network. Rather, the *physical* dynamical system – the entity at the physical level that corresponds to the mathematical dynamical system – is the network's *overall physical diachronic-potentiality profile*: its total dispositional profile of potential temporal transitions from one total activation-state to another. Thus, what corresponds physically to high-dimensional relative-position relations among representational points are the corresponding relations, within the physical-disposition profile, among the various total-activation states that realize TCS's. In short, non-intrinsic mathematical structure in the representational system is physically embodied in non-intrinsic physical-disposition structure. In principle, such mathematical and physical structure is much richer than the kinds of structure to which classicism resorts – viz., intrinsic mathematical structure, embodied in intrinsic physical structure.

Now consider the cognitive/mathematical realization relation, which pairs TCS's with points on the high-dimensional activation landscape. Since structure in the representational system is a matter of relative-position relations among representational points, an appropriate realization relation will position these points on the activation landscape in a highly non-arbitrary, highly systematic way – a way that works hand in glove with the landscape topography itself to yield temporal transitions from one representational point to another that are highly and systematically content-appropriate.[19]

In classicism, cognitive/mathematical realization is typically construed as a fairly straightforward, fairly uniform, relation: the idea is that intrinsic formal-syntactic structure of representations is recursively specifiable, and that semantic content can be directly gleaned

[19] Certain connectionist models like those of Pollack (1990) and Berg (1992) employ a training technique called the "moving target strategy" in which the representations change along with the weights in a process of controlled co-evolution. The network gets successively altered at the physical level via successive changes in the connection weights, with the network's current tentative internal representations continually figuring as "training targets" in the next stage of learning. With sufficient training the system converges on an appropriate combination of weights and representations. At higher levels of description this means that there is a progressive co-evolution of two interrelated factors. On the one hand, the dynamical system itself is being altered by weight changes; the local topography throughout the high-dimensional activation landscape is being progressively *molded*. On the other hand, because of the moving target strategy, there is also progressive refinement of the realization relation from intentional states to points on the activation landscape.

from intrinsic formal-syntactic structure by means of a translation manual that recursively specifies the compositional semantics of the representations. The realization relation between the atomic semantic constituents of thought-contents and the formal-syntactic atoms of representations is typically arbitrary; the recursion picks up from there.

The DC framework, by contrast, envisions a much less uniform realization relation. Since the activation landscape of a suitable neural network will be enormously complex and non-homogenous, and since human cognitive transitions are evidently too subtle and complex to conform to a tractably computable transition function, the cognitive/mathematical realization relation is likely to be very complex too. Also, since the core idea is that the shape of the activation landscape and the positioning of representational points on the landscape are "made for each other" under Nature's engineering design for human cognition, the realization relation will largely lack the kind of arbitrariness it has in classicism.

So the DC framework invokes a potentially more powerful kind of mathematics at the functional-mathematical level of description, and envisions a highly complex – but also highly systematic – cognitive/mathematical realization relation. The overarching conception of Nature's cognitive engineering is this: the neural network's activation landscape is so shaped, and representational points are so positioned on it, that temporal trajectories along the landscape from one representational point to another accord with a highly content-appropriate transition function over total cognitive states – a function so complex and subtle that it is not tractably computable. The DC framework thus jettisons assumptions (3) and (5) of classicism: that human cognition conforms to a tractably computable transition function over cognitive states, and is subserved mathematically be representation-level computation.

In general, the failure of a cognitive transition function (a CTF) to be tractably computable could be the result of either or both of two factors: (1) the dynamical system itself, whose mathematical state transitions might not be tractably computable; or (ii) the way TCS's are realized as points in the dynamical system's state space. Concerning the first factor, dynamical systems whose transitions are computable are actually a relative rarity, and it is certainly possible for noncomputable dynamical systems to be subserved by neural networks – at least if the networks are made more analog in nature by letting the nodes take on a continuous range of activation values, and/or letting them update themselves instantaneously rather than by discrete time steps.

Moreover, even if the mathematical state-transitions of the dynamical system are tractably computable (as they are for current connectionist systems, which are usually simulated on standard computers), the CTF subserved by the dynamical system might fail to be tractably

computable anyway, because realization may not be tractably computable. Consider the function that maps points in a dynamical system's state space to TCS's. (This is the converse of the realization relation; call it the *realizes-function*.) This function itself need not be tractably computable; on the contrary, the realizes-function envisioned by the DC framework is likely to be too complex to be tractably computable. If it isn't, then obviously one could not compute cognitive transitions in the following manner: (i) start from a TCS, (ii) find a point in state space that realizes that TCS, (iii) compute the dynamical system's trajectory through state space, and (iv) for each point p on the trajectory, compute the TCS (if any) realized by p. If the realizes-function is not tractably computable, then step (iv) will not be possible in general. (Also, in general there need not be any way to perform step (ii) by tractable computation either.) If the cognitive transitions subserved by the dynamical system are not computable in *this* way, they need not be tractably computable in any other way either. Thus, there can be a network-subservable dynamical system S, a CTF C, and a realizes-function R such that

(i) S's mathematical state-transitions are tractably computable;
(ii) S subserves C, under R; and yet
(iii) C is not tractably computable.

So underlying (tractable) computability at the mathematical level does not imply (tractable) computability at the cognitive level. One should not infer, from the fact that a system's mathematical state transitions are tractably computable, that it cannot subserve a CTF which fails to be tractably computable. Nor should one suppose that an algorithm for computing the system's mathematical state-transitions automatically counts as an algorithm for computing its cognitive transitions, i.e., as a set of PRL rules.[20]

The DC framework, as Tienson and I have articulated it, also incorporates a variety of further ideas about cognitive engineering, both from classicism and from connectionism. Among the principles adapted from classicism are these: (1) Representations exhibit *syntactic structure*, although syntactic constituency is construed in terms of relative-position relations in activation space among representational points, rather than (as in classicism) in terms of the intrinsic formal-mathematical structure

[20] Remember: PRL rules refer solely to the structural aspects of representations that play representational roles themselves. Rules for computing a network's activation transitions do not necessarily meet this condition – especially if all representations are "distributed," so that activation values of individual nodes have no representational content by themselves.

of representations.[21] (2) Representations are *productive*: there are vastly many cognitive states realized by points on the landscape, far more cognitive states than the cognitive system will ever actually instantiate in its lifetime; also, when a new representation is acquired in learning, realizations of vastly many complex representations involving this new content are automatically determined, throughout the activation landscape. (3) Representations exhibit *systematicity*, of the sort emphasized by Fodor and Pylyshyn (1988): if the cognitive system can represent that aRb (i.e., individual a bears relation R to individual b) and can represent that cRd, then it can also represent that aRd and that cRb.

Ideas from connectionism are incorporated into the framework too, and are suggestive of possible ways of bypassing persistent problems in computational cognitive science (like the frame problem). Among the principles adapted from connectionism are these: (1) Cognitive systems contain very rich morphological content: information that is accommodated during certain cognitive trajectories on the activation landscape without being actively (occurrently) represented by any point along the trajectory, and is mathematically embodied (given the overall relative positioning of representational points on the landscape) in the intricate high-dimensional topography of the landscape itself. Morphological content typically figures heavily in central cognitive processes like belief fixation. (2) Cognitive systems learn new information not by storing explicit representations in memory repositories, but rather by being "trained up" via weight changes which simultaneously mold the activation landscape and refine the cognitive/mathematical realization relation; these weight changes induce morphological embodiment of the information, suitably integrated with other morphological information. (3) Cognitive systems remember things not by finding and fetching explicit representations from mental file cabinets, but rather by automatically re-creating representations that are appropriate both to the system's current representational state (including information from sensory inputs) and to relevant morphological content.

The DC framework, like the other foundational approaches in cognitive science, is a "just-so story" about the mind: at present the quality and quantity of evidential support for this general view of mentality is still fairly meager.[22] Still, those inclined to think that human cognition

[21] This is a natural mathematical construal of the kinds of nonclassical syntactic constituency exhibited by Paul Smolensky's tensor product representations (Smolensky 1990, in press; Legendre, Miyata, and Smolensky 1991) and by Jordan Pollack's recursive auto-associative representations (Pollack 1990, Berg 1992). See Section 5.2 of Horgan and Tienson (1994), and Sections 5.1 and 9.3 of Horgan and Tienson (1996).

[22] Some of the connectionist work that Tienson and I have found especially suggestive and relevant is in the papers cited in Note 21.

is too complex and too supple to be subserved by representation-level computation need not retreat into silence or mysticism, because now there is a non-computationalist game in town.

Agnosticism about the various competing approaches is one epistemically reasonable stance, given the current evidential situation. Insofar as one is inclined to place one's epistemic bets, however, the persistent, in-principle-looking problems encountered by classicism suggest that the smart money is on the Dynamical Cognition framework.[23]

Department of Philosophy
University of Memphis
Memphis, TN 38152
USA

References

Amit, J. (1989). *Modeling Brain Function: The World of Attractor Neural Networks*. New York: Cambridge University Press.

Bechtel, W. (1991). "Connectionism and the Philosophy of Mind: An Overview." In Horgan and Tienson (1991).

—— (1993). "The Case for Connectionism," *Philosophical Studies* 71, 119–54.

—— and Abrahamsen, A. (1991). *Connectionism and the Mind: An Introduction to Parallel Processing in Networks*. Cambridge, MA: Basil Blackwell.

Berg, G. (1992). "A Connectionist Parser with Recursive Sentence Structure and Lexical Disambiguation." In AIII-92: *Proceedings of the Tenth National Conference on Artificial Intelligence*. Cambridge, MA: MIT Press.

Brooks, R. (1991). "Intelligence without Representation," *Artificial Intelligence* 44, 139–59.

Churchland, P. M. (1988). *Matter and Consciousness*, revised edition. Cambridge: MIT Press. Pp. 146–55.

—— (1989). *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*. Cambridge, MA: MIT Press.

[23] This paper is based on an overview talk on connectionism at the 1994 Eastern Division meeting of the American Philosophical Association, also presented to cognitive science groups at the University of Alabama at Birmingham and at the University of Mexico. I thank Bill Bechtel (the commentator at the APA session), David Chalmers, George Graham, Brian McLaughlin, John Tienson, and Mark Timmons for comments and discussion. The paper draws extensively on work published jointly with Tienson, and the ideas expressed are his as much as mine.

—— (1995). *The Engine of Reason, the Seat of the Soul: A Philosophical Journey into the Brain*. Cambridge, MA: MIT Press.
—— (1986). *Neurophilosophy: Toward a Unified Theory of the Mind/Brain*. Cambridge, MA: MIT Press.
Clark, A. (1993). *Associative Engines: Connectionism, Concepts, and Representational Change*. Cambridge, MA: MIT Press.
Clark, A. and Toribo, J. (1994). "Doing without Representing?" *Synthese* 101, 401–31.
Davies, M. (1991). "Concepts, Connectionism, and the Language of Thought." In W. Ramsey, S. Stich, and D. Rumelhart (eds.), *Philosophy and Connectionist Theory*. Hillsdale: Lawrence Earlbaum Associates.
Dreyfus, H. and Dreyfus, S. (1990). "Making a Mind versus Modeling the Brain: Artificial Intelligence Back at a Branch-Point." In M. Boden (ed.), *The Philosophy of Artificial Intelligence*. New York: Oxford University Press.
Fodor, J. (1983). *The Modularity of Mind: An Essay on Faculty Psychology*. Cambridge, MA: MIT Press.
Fodor, J. and Pylyshyn, Z. (1988). "Connectionism and Cognitive Architecture." In S. Pinker and J. Mehler (eds.), *Connections and Symbols*. Cambridge, MA: MIT Press.
Freeman, W. (1991). "The Physiology of Perception," *Scientific American* 264, no. 2, 78–84.
Gleick, J. (1987). *Chaos: The Making of a New Science*. New York: Viking.
Horgan, T. (1993). "The Austere Ideology of Folk Psychology," *Mind & Language* 8, 281–97.
Horgan, T. and Graham, G. (1991). "In Defense of Southern Fundamentalism," *Philosophical Studies* 62, 107–34.
Horgan, T. and Tienson, J. (1988). "Settling into a New Paradigm," *Southern Journal of Philosophy* 24, Spindel Conference Supplement, 97–113. Reprinted in Horgan and Tienson (1991).
—— (1989). "Representations without Rules," *Philosophical Topics* 17, 27–43.
—— eds. (1991). *Connectionism and the Philosophy of Mind*. Dordrecht: Kluwer.
—— (1992). "Structured Representations in Connectionist Systems?" In S. Davies (ed.), *Connectionism: Theory and Practice*. New York: Oxford University Press.
—— (1994). "A Nonclassical Framework for Cognitive Science," *Synthese* 101, 305–45.
—— (1995). "Connectionism and the Commitments of Folk Psychology," *Philosophical Perspectives* 9, 127–52.
—— (1996). *Connectionism and the Philosophy of Psychology*. Cambridge, MA: MIT Press.

Hume, D. (1978). *Treatise of Human Nature*. Oxford: Oxford University Press.

Legendre, G., Miyata, Y., and Smolensky, P. (1991). "Distributed Recursive Structure Processing." In D. Touretzky and R. Lippman (eds.), *Advances in Neural Information Processing* 3. San Mateo: Morgan Kaufmann.

Marr, D. (1982). *Vision*. New York: W. H. Freeman.

McLaughlin, B. (1993). "The Connectionism/Classicism Battle to Win Souls," *Philosophical Studies* 71, 163–90.

Pollack, J. (1990). "Recursive Distributed Representations," *Artificial Intelligence* 46, 77–105.

Ramsey, W., Stich, S., and Garon, J. (1990). "Connectionism, Eliminativism, and the Future of Folk Psychology," *Philosophical Perspectives* 4. Reprinted in Ramsey, Stich, and Rumelhart (1991).

Ramsey, W., Stich, S., and Rumelhart, D., eds. (1991). *Philosophy and Connectionist Theory*, Hillsdale: Lawrence Earlbaum Associates.

Sejnowski, T. and Rosenberg, C. (1987). "Parallel Networks that Learn to Pronounce English Text," *Complex Systems* 1, 145–68.

Skarda, C. and Freeman, W. (1987). "How Brains Make Chaos in Order to Make Sense of the World," *Behavioural and Brain Sciences* 10, 161–95.

Smolensky, P. (1990). "Tensor Product Variable Binding and the Representation of Symbolic Structures in Connectionist Systems," *Artificial Intelligence* 46, 159–216.

—— (in press). "Constituent Structure and Explanation in an Integrated Connectionist/Symbolic Cognitive Architecture." In C. and G. MacDonald (eds.), *The Philosophy of Psychology: Debates on Psychological Explanation*. Cambridge, MA: Basil Blackwell.

Stewart, I. (1989). *Does God Play Dice? The Mathematics of Chaos*. Cambridge, MA: Basil Blackwell.

Tienson, J. (1988). "An Introduction to Connectionism," *Southern Journal of Philosophy*, Spindel Conference Supplement on Connectionism and the Philosophy of Mind. Reprinted in J. Garfield (ed.), *Foundations of Cognitive Science: The Essential Readings*. New York: Paragon House.