Putting Together Connectionism - again *

Paul Smolensky

CU-CS-378-87

University of Colorado at Boulder
DEPARTMENT OF COMPUTER SCIENCE

* This research has been supported by NSF grant IST-8609599 and by the Department of Computer Science and Institute of Cognitive Science at the University of Colorado at Boulder.

Putting Together Connectionism—again

Response to Commentary on *BBS* target article: On the Proper Treatment of Connectionism

Paul Smolensky

CU-CS-378-87 October, 1987

Department of Computer Science &
Institute of Cognitive Science
University of Colorado
Boulder, CO 80309–0430
(303) 492-8991
smolensky@boulder.colorado.edu

To appear in *The Behavioral and Brain Sciences* (1988)

This research has been supported by NSF grant IST-8609599 and by the Department of Computer Science and Institute of Cognitive Science at the University of Colorado at Boulder.

Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author and do not necessarily reflect the views of the National Science Foundation.

Table of Contents

If μ implements M, then this constitutes the strongest possible sense in which μ and M could be both valid descriptions of the same system S. If we take μ to be a connectionist account and M a symbolic account, then assuming that μ is an implementation of M is the view of connectionism I will call *implementationalist.* The implementationalist view is rejected by PTC. This rejection is stated in (8c), and the wording of (8c) is designed precisely to reflect the characterization of the implementation relation given in the preceding paragraph.

If μ is not an implementation of M, another obvious possible relation between μ and M is that they have nothing principled to do with each other. If S is a system that is described at level L_μ by μ, then there is no description at any level of S that bears any significant similarity to M, except possibly for isolated accidental coincidences. In this case, M can have no role to play in explaining the behavior of S.[2]

If μ is a connectionist account and M is a symbolic account, this relation corresponds to the *eliminativist* position: connectionist accounts eliminate symbolic ones from cognitive science. Like the implementationalist position, the eliminativist position is also rejected by PTC.

Table A presents the implementationalist and eliminativist positions, along with a number of other relevant positions on the relation between connectionist and symbolic accounts. All positions in Table A assume some degree of validity of the connectionist approach; they differ in their assessment of the validity of the symbolic approach and the relation between the two approaches.

Insert Table A about here

As Table A indicates, and as I will shortly discuss, PTC adopts a view in some sense intermediate between the far-left eliminitivist and far-right implementationalist views. The intermediacy of the PTC position allows us to understand an interesting phenomenon that occured in the commentaries. A commentator leaning toward one or the other of the extreme views correctly saw in PTC a rejection of their view. Their response was to conclude that PTC embraced the other extreme, and to launch on PTC their favorite attacks on the opposite extreme. Thus we see why **Touretzky**—whose connectionist models (eg., Touretzky 1986; Touretzky & Hinton 1985) probably come closest to realizing the implementationalist strategy—identifies the contribution of the target article as "defin[ing] the eliminative stance" at the same time that **Hanson** calls PTC a "'strong implementational' view of connectionism." Implicitly, the logic in these commentaries stems from the following assumption:

(EF) **The Extremist Fallacy:**
There exist only two viable views on the connectionist/symbolic relation: eliminativism and implementationalism. Any approach that clearly rejects one view must either embrace the other or be incoherent.

Some commentators, seeing correctly that PTC rejects *both* extreme positions, followed (EF) to the conclusion that PTC is incoherent (eg., **Antony & Levine** and **Dietrich & Fields**).

It therefore becomes crucial to establish that (EF) is indeed a fallacy; that there is a coherent perspective that rejects both extremes. The target article is, of course, intended to argue for just this conclusion. The article summarizes a program of research carried out in the intermediate perspective, to illustrate: that the framework is viable, that it can lead to interesting research; and that it has the potential to account for a variety of aspects of cognition exceeding that which either extreme view can handle separately.

Another argument is more hinted at than formally presented: an argument by analogy with physics, in which the intermediate position of PTC is likened to the relation between the microphysics of the quantum theory of matter and the macrophysics of Newtonian mechanics. Note that the point of the analogy is to show by illustration that an intermediate view like that of PTC cannot be simply dismissed as intrinsically incoherent, as in (EF).

[2] The case of accidental coincidences is what **Woodfield & Morton** call "killing two birds with one stone."

The format of this response is indicated in the following table of contents, which lists, in order of appearance, the commentaries considered (sometimes in footnotes).

1. Levels of analysis

The major issue discussed in the target article was levels analysis employed by various approaches to cognitive science. Several of the commentaries misconstrued the PTC position on this issue, either explicitly or implicitly. The target article focussed on what the PTC account *is*; I will now devote more attention to what it *is not*, providing a better framework within which to respond to the commentaries.

1.1. A framework for discussion

Suppose we are given two computational accounts at different levels; call the lower- or "micro"-level description μ, and the higher- or "macro"-level description M. μ might be an assembly-language program, or the differential equations describing the circuits in a von Neumann computer, or the differential equations describing activation passing and connection strength modification in a connectionist network. M might be a Pascal program, or an OPS-5 production system for solving arithmetic problems. The question is: What possible relations might hold between μ and M? (The following discussion expands upon that of Pinker & Prince, 1987).

The first possibility is the most straightforward: μ is an implementation of M. The notion of implementation is brought to us primarily by the von Neumann computer, and throughout this discussion I will take "implementation" to mean exactly what it means in that context.[1] For my purposes, the crucial aspect of the implementation relation is this. Suppose we have a physical system S which at some level of description L_μ is exactly performing the computation μ; that is, if we write down the laws governing the dynamics and interactions of those aspects of the system state that are characteristic of level L_μ, we find these processes to be exactly described by μ. If μ is an implementation of M, we are guaranteed the following: the states of this same system S have characteristics at a higher level L_M which evolve and interact exactly according to M: these characteristics define a description of S at the higher level L_M for which M is an *complete, formal, and precise* account of the system's computation.

[1] Some commentators use "implementation" loosely, apparently equating it with a weaker notion such as instantiation (eg., **Rey** and **Van Gulick**). In the present context, it is advisable to use terms for various relations between levels with precision; all statements about the subsymbolic approach not being "mere implementation" refer to this specific sense of "implementation."

The micro/macrophysics analogy was not construed uniformly by the commentators in the way it was intended, so let me expand upon it here. Some readers may have taken the comparison to Newtonian physics as deprecatory—quite the opposite of the intended reading. Newtonian mechanics was chosen as a case where a macrotheory is scientifically rock-solid, and explanatorily essential, despite the fact that it is known to be not literally instantiated in the world, according to the current best theory, the microtheory. As Stich points out, the fundamental elements in the ontology presumed by the macrophysics simply *cannot* literally exist according to the ontology of microphysics (rigid bodies, deterministic values for observables, Galilean invariance of physical laws, ...). In a strictly literal sense, if the microtheory is right, the macrotheory is wrong;[3] it is in this quite non-trivial sense that I describe the micro- and macro-theories as "incompatible" (contrary to **Cleland, Dietrich & Fields**, and **Van Gulick**). It does *not* however follow that the macrotheory is *explanatorily irrelevant*: in the world of real explanations, Newtonian explanations are (at least) as important as quantum ones (within their proper domain). The position on explanation that PTC relies on goes something like this:

(AE) **The Principle of Approximate Explanation:**
Suppose it is a logical consequence of a microtheory that, within a certain range of circumstances, C, laws of a macrotheory are valid *to a certain degree of approximation.* Then the microtheory licenses *approximate* explanations via the macrotheory for phenomena occuring in C. In very special cases, these phenomena may admit more exact explanations that rest directly on the laws of the microtheory (without invoking the macrotheory), but this is not to be expected generically: for most phenomena in C, *the only available explanation will be the (approximate) one provided by the macrotheory.*[4]

This principle illustrates a third relation that can exist between a micro-account μ and a macro-account M: M *approximately* describes the higher level behavior of S, not *accidentally* but because there are systematic, explanatorily-relevant relationships between the computations performed by μ and M. That the relationships between μ and M are "systematic" manifests itself in principle (AE) through the proofs (or less rigorous logical arguments) that, given that the laws μ hold at the micro-level, it follows that the laws of M hold at the macro-level. Let us call this relationship between μ and M *refinement*: μ is a refinement of M.

Refinement, not implementation, is the relation between micro- and macrophysics. Figuratively speaking, "programs" written in Newtonian physics that depend on strict determinism or absolute simultaneity will not "run" correctly in a world of quantal uncertainties and Einsteinian relativities. If quantum theory were an implementation of Newtonian mechanics, it would be guaranteed that any phenomenon describable in the Newtonian vocabulary would be governed *exactly* by Newtonian laws; quantum theory would be needed only for microevents not describable at the higher level of Newtonian theory. It is just such a guarantee that ensures that a program written in Common Lisp will provide an exact higher level description of any computer running that program on top of a genuine implementation of Common Lisp.

It is useful to be a bit more concrete about one sense in which the macrotheory approximates the microtheory. In physics, the passage from the microtheory to the macrotheory is a certain *limit* in which various parameters of the system being described approach extreme values. (For example, Newtonian mechanics corresponds to a limit of relativistic quantum theory in which, loosely speaking, masses of bodies approach infinity and speeds approach

[3] Note that just the reverse is true of implementations: there, according to the micro-account, the ontology of the macro-account *must* exist, since it can be logically and exactly derived from the micro-account. If the microtheory is right, the macrotheory *must* be right.

[4] That the macrotheory has explanatory priority for most phenomena in C seems to be behind comments of **Cleland, Woodfield & Morton**, and **Pinker & Prince**. Given this, the hypothesis in (10) of the target article, that the subsymbolic account is complete, should be construed not to refer to *explanatory* completeness; rather the sense of "completeness" in which quantum theory *in principle* applies to all phenomena, while Newtonian mechanics does not.

zero.) Thus the mathematical analysis of the emergence of the macrotheory from the microtheory involves taking limits in which certain idealizations become valid. In the cognitive case, there are many limits involved in the passage from subsymbolic models to symbolic models. Among these limits are: number of connectionist units or strengths of connections approach infinity (allowing "soft" properties to become "hard," and allowing memory capacity limits to be idealized away for "competence" theory), relaxation or learning time approaches infinity (allowing, eg., stochastic inference or learning to converge to theoretically known limits), and overlap (inner product) of vectors stored in memory approaches zero (orthogonality: eliminating interference of items of memory).

Since the formal relationship between the micro- and macro-levels presumed here is one of convergence in the limit, the PTC position in Table A has been called *limitivist.* This name is also appropriate in that the microtheory explicitly *limits* the applicability of the macrotheory to certain circumstances C, and specifies the *limits* of its accuracy within C.

Having hit the main points of Table A, let's run through the positions outlined in the Table, from the extreme left to the extreme right.

There are two brands of eliminitivists indicated: the furthest left position maintains that the science of cognition will entail accounts at the neural level, and that higher levels, whether they be those of the symbolic or subsymbolic paradigm, can furnish no more than folklore. The only scientifically valid cognitive models are real neural models. Slightly less left is the position that connectionist models offer scientifically valid accounts, even if they are at some level higher than the neural level, but accounts at levels higher than that of connectionist nodes and links, including symbolic models, have no scientific standing.

Left of center is the position taken by PTC, that accounts at the neural, subconceptual and conceptual levels can all provide scientific explanations. The conceptual level offers explanations that are scientifically valid provided due consideration is taken of their approximate and restricted nature; the exactness and coverage of cognitive phenomena afforded by the subconceptual accounts is much greater. Within this view, symbolic methods cannot provide complete, formal, and precise accounts of intuitive processing (8c), but this leaves a number of important roles for symbolic accounts (briefly mentioned in the target article following 8c):

- describing consciously mediated (non-intuitive) processes—including many of the phenomena so important to philosophers, such as conscious reasoning;
- describing isolated aspects (i.e., not "complete accounts") of performance;
- giving general (as opposed to detailed and "formal") ways of understanding and thinking about cognitive processes.
- describing intuitive competences: abstractions away from performance (i.e., not "precise"), eg., in language processing (contrary to **Rey**, valid competence theories are consistent with PTC);

Right of center is the *revisionist* position, which sees as a primary function of connectionist theory the revising of symbolic theory; after such revision, this view has it, symbolic theory will provide an complete, formal, and precise account of cognition at the macro-level. A favorite way to imagine the revisionist scenario playing out is to modify symbolic theory by relegating certain processes to connectionist networks: eg. perception, memory, production matching, and other "low level" operations. The image that emerges is that the mind is a symbolic computing engine with handy connectionist peripherals to which it can farm out certain low-level chores that connectionist nets happen to have a talent for. Since, as we all know, the left half of the brain does hard, rational symbolic-ey processing while the right half does soft, squishy, connectionist-ey processing,[5] this version of the revisionist story sees the mind as a house divided, right and left working side-by-side despite their profound differences. Daniel Andler and I call this arrangement by its French name, *cohabitation.* (**Woodfield & Morton** call this "division of labor.")

[5] It's obvious the two sides were named by someone looking from the wrong direction.

A more subtle, but vaguer, revisionist view envisions revision of the way the basic machinery of symbolic computation is used in cognitive models, based on the way symbolic operations are actually realized in the connectionist substrate (Pinker & Prince, 1987). I am not aware of any suggestions for how this might be carried out in practice.

One difference between a PTC and revisionist view can be illustrated through the commentary of **Lloyd**: he clearly places priority on higher, conceptual-level accounts of cognition, he resists PTC's move to give theoretical parity (or even priority) to the lower, subconceptual level, and he looks to connectionism primarily as a way of developing new, better, formal accounts at the conceptual level. Viewed through PTC's Newtonian/quantum analogy, Lloyd's view becomes: "What we really want out of physics is the study of macroscopic, everyday, rigid bodies; quantum mechanics should be used primarily to provide us with better theories of such bodies, and not to shift our attention to lower levels that are not properly the study of physics."

The final, far-right view is the implementationalist view already discussed.

The target article stated that PTC rejects "blandly ecumenical" views; this term was intended to cover both the *cohabitation* version of revisionism and implementationalism. The sense in which these views are "bland" is that they involve no reconstruction of the core of the cognitive architecture presumed by the symbolic approach; they just involve realizing low-level operations in connectionist terms. In the *cohabitation* approach, selected low-level processes in the architecture get done with connectionist networks, giving a higher level performance that is potentially different from the symbolic components they replace (eg., the new memory is content-addressable whereas the old was not). In the implementationalist approach, connectionist networks perform the primitive operations needed to support all of symbol processing, but they do it in such a way that viewed from the higher level, the computations are the same as they were before.

By contrast, the PTC approach requires a complete reconstruction of the cognitive architecture. It doesn't recycle the symbolic core, adding connectionist peripherals or providing connectionist implementations of all the Lisp primitives. PTC is "incompatible" with the symbolic approach because it *does* involve reconstructing the cognitive architecture. PTC is self-consciously ecumenical—but not blandly so.

To appreciate this point, it is helpful to draw out the methodological import of the distinction between the PTC and blandly ecumenical views. Consider a core cognitive function, say, language comprehension. Given the three views, *cohabition*, implementational, and PTC, what is the job of the connectionist researcher? A researcher of the *cohabitation* school takes a symbolic program for language comprehension and asks, "How can I rewrite this program to make use of a connectionist memory and connectionist best-match routine?" An implementationalist asks, "What are the primitive symbolic operations into which this program compiles, and how can I build connectionist nets to implement them?" The PTC approach spawns several questions: "What aspects of human performance are being captured by this program, and which are being missed? What are the computational abstractions being used in the program that allow it to capture what it captures? What are natural ways of instantiating those abstractions in connectionist computation? Are there ways to use connectionist computation to model the aspects of human performance that the symbolic program is missing?"

These differences between the methodological implications of the implementationalist and PTC views are illustrated in Figure A. At the top level are information-processing abstractions such as memory, constituent structures, attention, and so forth. At the next level are the particular formal instantiations of these abstractions that appear in symbolic cognitive science. Below these on the right branch are connectionist implementations of these symbolic computational elements. This is the implementationalist branch. On the left branch, the high-level abstractions have been instantiated directly in their natural PTC-connectionist form, without passing through the symbolic formalism. But because these PTC-connectionist instantiations are reifying the same kinds of abstractions, and because the symbolic formalism does capture important aspects of human cognition, there is a relation between the connectionist instantiations on the left branch and the symbolic instantiations on the right branch: the former are a refinement of the latter, i.e., the symbolic formalism is an approximate higher level description of the PTC-connectionist formalism (as opposed to an exact higher level description of the implementationalist-connectionist formalism on the right). Again, the main point is this: the right, implementationalist, branch preserves the symbolic cognitive architecture, while the left, PTC branch requires a reconstruction of the cognitive architecture in which the basic computational abstractions acquire new, inequivalent, instantiations.

Insert Figure A about here

As the last two paragraphs show, there are important methodological implications for actual connectionist modeling of whether connectionist models literally *implement* symbolic models, or whether the two kinds of models merely *instantiate common underlying principles*. Thus it is important to ask (with **Van Gulick**), "at what level will we find powerful insightful cognitive generalizations"—but it is *also* important to ask (contrary to **Van Gulick**) "at what level are complete precise formal cognitive descriptions to be found," for it is this question that determines which branch of Figure A we are to follow.

Chandrasekaran, Goel, & Allemang are right to emphasize the importance of "information processing abstractions"; these are the elements at the top level of Figure A. But it is also necessary to emphasize the importance of the particular shape these abstractions take when they are formalized in a particular framework.

1.2. Commentaries compatible with PTC

Having placed PTC explicitly in the context of alternative views of connectionism, I now proceed to direct replies to commentary, starting with those consistent with the PTC view.

Hofstadter's commentary illustrates his view of conceptual-level interactions, a view that seems to cry out for instantiation of concepts as something computationally akin to patterns of activity: patterns that take context-dependent forms, overlap with a rich topology, and support subtle conceptual-level interactions that emerge from simpler interactions between the elements of the rich internal structure of these concepts. The fluid conceptual interactions of common sense demanded by Hofstadter are, on the PTC account, built into the very fabric of the architecture: they are not add-ons to an otherwise brittle system. Hofstadter's view is not only very close to PTC, it is one of PTC's chief sources. Many of the elements of PTC have rather direct counterparts in the writings of Hofstadter: subsymbols (Hofstadter 1985, p. 662), the subconceptual level hypothesis (Hofstadter 1985), the relation between conceptual, subconceptual, and neural models (Hofstadter 1979, 569–573), symbols and context dependence (Hofstadter 1979, 349–50), and even computational temperature (Hofstadter 1983). While Hofstadter has articulated these principles, and argued extensively for them, he has incorporated them into research limited to the conceptual level; the methodological conclusion that PTC draws is, of course, quite different.

Dellarosa's commentary raises the issue of whether connectionist processing should be viewed as "association" or "inference." ("Associationism" is also raised—but as an ominous accusation—by **Lindsay**.) The view favored in the target article, and pushed even further by **Golden**, is that the fundamental processing in connectionist networks is *statistical inference*, which sits somewhere intermediate between the notions of "pure association" and logical inference. Since "pure association" is an undefined, informal notion, it is difficult to say in which respects the processes underlying the most powerful connectionist models exceed that of pure association. But the kind of statistical inference underlying the harmony model of circuit reasoning, discussed in Section 9.2 of the target article, seems more powerful than "mere" association, in its ability, in the appropriate limit, to give rise to a competence that is correctly characterized through logical inference. It is probably best to say that just as predicate calculus is Aristotle's notion of inference dressed up and gone to college, so the statistical inference of connectionist networks is Humean association with a Master's degree.

Rueckl makes the important point that softened conceptual-level formalisms such as fuzzy logic can be used to try to formalize subsymbolic models at the conceptual level, but they will in general fail because they do not capture enough of the internal structure of concepts to be able to account for the causal interactions of those concepts. I might add a technical note: Rueckl is correct in stating that it's not possible to predict much if all that's known is the degree to which a pattern is present, but much more can be predicted if instead what's known is the degree to which a pattern overlaps a complete set of patterns. This is in fact the basis of the conceptual-level analysis of Smolensky (1986b), which is summarized in Section 9.3 of the target article.

Most of **Dyer**'s comments seem to be consistent with the PTC position, and I have nothing substantial to dispute in, or add to, his observations.

1.3. Misunderstandings of the PTC position

The framework presented above allows us to clear up confusions about the PTC position present in a number of commentaries.

Both **Harnad** and **Quarton** take my use of "symbolic" and "subsymbolic" to refer to levels. In fact, these terms refer to paradigms for cognitive modeling, *not* levels. Harnad is right to question whether "subsymbolic" refers to a lower level than "symbolic": it doesn't. The symbolic and subsymbolic paradigm, as defined in the target article, are approaches to cognitive modeling that use, respectively, symbolic and subsymbolic models, each of which can be analyzed at various levels of analysis. As Table 2 illustrates, the symbolic/subsymbolic distinction is orthogonal to the level distinction between conceptual and subconceptual. These levels are semantic levels: they refer to mappings between formal models and what they represent. On the side of what is represented, the conceptual level is populated by consciously accessible concepts, while the subconceptual level is comprised of fine-grained entities beneath the level of conscious concepts. On the side of the formal models, for connectionist models the conceptual level consists of patterns of activity over many units and their interactions, while the subconceptual level consists of individual units and their interconnections. For symbolic models, the conceptual level consists of symbols and the operations that manipulate them, and lower levels (no one of which has the distinction of being singled out as "the subconceptual level") consist of the finer grained operations on which symbol manipulation is built.

In other words, the level distinctions involve levels of aggregation, what **Harnad** calls the "molar/molecular or macro/microlevels of description," just as in the case of macro/microphysics, the basic analogy that was provided for understanding the intended sense of "levels." (**Lakoff** further distinguishes this use of "levels" from a related but different usage in linguistics.)

As in Table 2, **Quarton**'s two-dimensional array of models illustrates the orthogonality of levels of description and models being described. His commentary is quite helpful, and usefully distinguishes "simulation relevant" and "simulation irrelevant" lower levels. The commentary unfortunately ignores, however, the crucial fact that different levels can be related in ways other than implementation; his picture handles levels in computer systems but cannot really accommodate the relevant relationship for PTC: the sense in which macrophysics is a higher level description of microphysics. What is needed is a vertical relation other than implementation, or, if models related by other than implementation are to be separated horizontally, an analysis of horizontal relations.

Some commentators found the thrust of the target article inconsistent with their construal of descriptive terms such as "incompatible," "inconsistent," and "blandly ecumenical." Rather than letting their understanding of the gist of the article guide them to interpretations of these unimportant terms that would lead to an overall consistent reading of the article, they prefered to stick with some a priori favorite characterization of these terms and get confused by imagined inconsistency.

For example, **Dietrich & Fields** seem to have grasped entirely the intent of PTC's limitivist position, yet because they did not see this position as representing "incompatibility" between connectionist and symbolic accounts, they prefered to see inconsistency. As explained above, there is a perfectly reasonable sense in which Newtonian mechanics and quantum theory are "incompatible"; in fact, this sense of incompatibility is sufficient to lead **Stich** to conclude that the microtheory can eliminate the scientific standing of the macrotheory.

Dietrich & Fields pursue their misconstrual of "incompatibility" to the conclusion that PTC must be committed to the lack of a consistent mapping between patterns of activity and concepts—for otherwise their would be in principle a "complete, formal, and precise" PTC account at the conceptual level. Here they ignore the word *tractable* in (8c). That such conceptual level accounts exist *in principle* is not the issue; the question is whether such accounts exist in sufficiently tractable form to serve the scientific needs of building models, making predictions, and providing explanations. (Besides, that the pattern-of-activity-to-concept mapping is imprecise is exactly the content of Section 7.2.)

Dietrich & Fields' claim that models can be given any semantic interpretation at any level seems to indicate that they have in mind a profoundly different sense of "level" from that used in the target article. In claiming that one can interpret neurons as representing grandmothers they appear to be blurring the distinction between mapping a *single* neuron's state onto a representation of grandmother and mapping *collective* states of a population of neurons onto such a representation. If we replace "neuron" by "node in a subsymbolic connectionist network," then this distinction is *precisely* that between giving semantic interpretation at the subconceptual level and at the conceptual level. **Lakoff** spells this out quite clearly.

Touretzky asserts that the PTC position on his bouncing thermostat is the eliminitivist one, (iv); in fact, the PTC position would be to develop equations correctly accounting for the bouncing, and to derive mathematically the result that the higher level rule is approximately satisfied. It should be possible in fact to derive the limits of this approximation: the amount of time after crossing the setpoint that "performance noise" will obscure the thermostat's "real competence," and conditions under which the competence will fail to appear at all (eg., subjecting the thermostat to rapid temperature oscillations that prevent it from equilibrating).

As stated earlier, several implementationalist-leaning commentators mistook PTC for eliminitivist; in addition to **Touretzky**, these include **Rey** ("connectionism in the end ought to replace ...") and **Schneider** ("... paradigm shift laying waste to its predecessors").

Bechtel is worried about how the connectionist conscious rule processor can do its job without actually being implemented, and thus violating (8c). But (8c) only refers to the *intuitive* processor, so this problem is a simple misunderstanding. Indeed Section 6 of the target article is devoted to implementing the conscious rule interpreter in connectionist networks.

1.4. Arguments against PTC's relation to the symbolic approach

Several commentators argue for positions in Table A other than the PTC position.

Dyer and **Touretzky** emphasize the importance of symbolic processes (eg. variable binding) in performing complex information processing; the PTC view is in agreement: it is necessary "to extend the connectionist framework to naturally incorporate, without losing the virtues of connectionist computation, the ingredients essential to the power of symbolic computation" (Smolensky 1987). While arguments like those of Dyer and Touretzky emphasizing the importance of symbolic computation are often viewed as arguments *in favor* of implementationalist or revisionist views of connectionism, they are really arguments *against* the eliminative view; they are therefore quite compatible with the PTC view. It must be realized, however, that since a PTC view, following the left branch of Figure A, insists on *reinstantiating the basic ideas behind symbolic computation in a fully connectionist fashion*, and not merely implementing the standard instantiations of those ideas, we will not know for some time yet whether the PTC approach can adequately martial the power behind symbolic computation.

The commentators who seem to advocate revisionist positions include **Chandrasekaran, Goel, & Allemang** ("Connectionist architectures seem to be especially good in providing some basic functions ... Symbolic cognitive theories can take advantage of the availability of connectionist realization of these functions ..."), **Schneider** ("it would be better to identify the weaknesses and strengths of each and examine hybrid architectures"), and, most explicitly, **Lloyd.**

Stich argues for an eliminativist position, prefering to view symbolic theory in analogy with caloric theory rather than with Newtonian mechanics. The moral he wants to draw from his analogy is that the microtheory (kinetic theory) eliminated the macrotheory from science. It is a strange moral to draw, however: there is no more spectacular (and classic) example in science of a microtheory that vindicated and refined—rather than eliminated—a macrotheory than that of kinetic theory (statistical mechanics) and thermodynamics. Whatever may have been the fate of the particular stuff called "caloric fluid," the scientific standing of macrotheory in this area is not in doubt. It is the view that successful micro-theories always eliminate macro-theories from science that I refered to in the conclusion of the target article as "naive ... eliminative reductionism," and Stich is correct that the PTC view rejects the eliminative conclusion. (**Woodfield & Morton** correctly emphasize that one cannot be both "emergentist"—the PTC position—and eliminativist about symbolic processing.) Stich is quite right to point out that in the traditional paradigm symbols are reified—have a hard and stable existence—to an extent that is not likely to emerge from connectionist networks. But it seems to be typical that when a macrotheory is reduced to a microtheory, what was seen before as a reified, hard, and stable substance (eg., caloric fluid, rigid bodies) is now viewed as a much more abstract entity emerging from the interaction of lower level entities that are (for the time being) viewed as the reified and stable substrate. It is not surprising that symbols and symbol manipulation should suffer the same fate.

Woodfield & Morton propose that the relation of symbolic to connectionist accounts may be different from all those included in Table A: a relation analogous to that between entering into a contract and signing one's name. I am unable to see how this intriguing proposal might work. The causal powers of contracts are instantiated in the world through cognitive systems that recognize name-signing and act upon that recognition according. How can the causal powers of symbols be instantiated analogously, without some system that recognizes relevant subsymbolic activity and acts upon that recognition? Or is that exactly what is being proposed?

Chandrasekaran, Goel, & Allemang propose to characterize the levels issue in terms of Marr's computational/algorithmic/implementational analysis, and want to say that the symbolic/connectionist debate is clouded by looking at implementational levels instead of computational or algorithmic levels. The target article emphasizes the importance of looking at the higher level properties of connectionist systems, and this is one respect in which the PTC approach differs from much connectionist research that is focussed more on the lower level. Harmony theory, for example, can be viewed in the Marr framework quite well: there are two rather clearly identifiable theoretical accounts at what can be called the computational and algorithmic levels; simulating a harmony model brings in an implementation level as well (Smolensky 1986a). Chandrasekaran, Goel, & Allemang are correct to point out how understanding a model at the higher levels greatly promotes understanding of what is "really doing the work" in the model, and avoids confusion over irrelevant details. The Marr framework is useful for better understanding an *individual* computational model, whether it is symbolic or connectionist. Marr's framework relates to *within model* level relationships; it can't do, however, for the *between model* relation that PTC hypothesizes between symbolic and connectionist models—unless the framework is expanded to permit algorithmic-level accounts that only approximately instantiate a computational account. But here again, the Marr view of levels is best suited for level relations that are found in machines (Marr's example is an adding machine); if approximation/refinement is crucial, as it is for PTC, why not replace machine-based level analogies for one that does full justice to the notion of refinement, like the microphysics/macrophysics analogy?

Antony & Levine want to deny PTC its place on the spectrum of Table A, essentially by arguing for the Extremist Fallacy, which they state in their concluding paragraph. Their argument is basically that either connectionism denies that symbolic entities (eg., constituent-structured data and structure-sensitive operations) have explanatory roles—eliminativism—or connectionism admits that these entities have explanatory roles, and therefore that connectionist models implement symbolic entities. This implicitly denies the Principle of Approximate Explanation on which PTC rests. Section 7.2 is an attempt to briefly show not that constituent structure can be "read onto" connectionist networks, as Antony & Levine state, but that constituent structure *has an important role to play in explanations* (albeit approximate ones) of the high level behavior of connectionist systems. Section 7.2 is quite explicit about this: "the *approximate* equivalence of the '*coffee* vectors' across contexts plays a functional role in subsymbolic processing that is quite close to the role played by the *exact* equivalence of the *coffee* tokens across different contexts in a symbolic processing system." Antony & Levine offer no response, and their argument relies critically on refuting (or ignoring) this crucial point.

1.5. The neural level

Lloyd, Mortensen, Ruekl, and, to a lesser extent, **Bechtel**, all advocate the pursuit of increased coupling between the neural and subsymbolic modeling. My point here is really not to argue against such a pursuit, but rather to be clear about the current gap between neural and subsymbolic models, to recognize the independently valid role that each has to play in cognitive science, to admit that each has its own set of commitments and goals, and to be open to any of a number of outcomes. It may happen that the gap between the two kinds of models will shrivel away; but there are reasons to believe (see footnote 7 and preceding discussion in the target article) that in fact the opposite is now occuring. It may be that Lloyd's golden age picture will come to pass, or it may be (as argued by Stone, see Section 2 below) that instead of one level between the neural and conceptual levels we will need many levels.

den Uhl makes the important point that Table 1 is based on typical current connectionist models that consist of a single module; if future models involve multiple modules, some of the '–'s in the table will change to '+'s. Table 1 is deliberately chosen to reflect the current state of the art, and will need to be kept up to date. The real question is whether the modular structure of future connectionist models makes contact with the modular structure of the brain, or whether the models' architecture will be driven by computational considerations that turn out not to be deeply tied to the neural architecture.

2. Treatment of connectionist models

Several commentaries address perceived inadequacies in the target article's treatment of connectionist models; except where noted below, I am basically in agreement with the commentators.

Touretzky argues that connectionist models of complex processes will have to introduce persistent internal state, modular structure, and built in mechanisms for complex operations such as variable binding. **den Uhl** convincingly elaborates the call for modular structure. Both commentators argue that the kind of mathematics that has so far contributed nearly all the technical insights of connectionism, the continuous mathematics of dynamical systems, will not continue to play this central role as connectionist models increase in their structure and complexity. This conclusion may be correct, but it seems reasonable to take as the working hypothesis that the mathematics describing current connectionist networks, to be the modules of future systems, will have to contribute importantly to the analysis of the whole, even if other kinds of mathematics also come into play.

Schneider argues that getting symbolic processing done in a connectionist network requires specially crafted networks, and that specially designed attentional mechanisms are needed.

Golden wishes to elevate subsymbolic principles of rationality, based on statistical inference, to the defining characteristic of the subsymbolic paradigm. While I accept the centrality of statistical inference to the paradigm, and its role in characterizing rationality, it seems overly restrictive to exclude other computational processes in dynamical systems, eg., motor control, which have yet to be shown to fit within a statistical inference framework. Given the extent to which Golden has been able to extend the statistical inference analysis of harmony theory and the Boltzmann machine, it may however in fact turn out that essentially all of subsymbolic processing will eventually be seen to fall within the boundaries of statistical inference.

Stone's elegant commentary makes the following point: if the target article succeeds at legitimating the hypothesis of *one* level intermediate between the conceptual and neural levels, and in characterizing its relations to levels above and below, why not repeat the argument to legitimate numerous levels, determined pragmatically and perhaps domain-specifically in response to demands for explanation of various cognitive regularities? Put differently, the "subconceptual level" hypothesized by the target article can, in fact, be viewed as containing a number of sublevels, all lying between the conceptual and neural, all characterized by connectionist processing, more low level accounts being refinements of higher level ones. Sorting out this fine-structure can be expected to be a domain-specific enterprise.

Lakoff points out that if subsymbolic models do not remain in their current isolated status but are somehow tied down to neural systems in the body, then the semantics of patterns of activity are not free for the theorist to invent: they are automatically grounded by the organism. This seems an important philosophical point, but one that cannot really do any modeling work until the gap is bridged (at least partly) between the subconceptual and neural levels—unless Lakoff's research program can be pulled off: the grounding of subsymbolic models in the image schemas of cognitive semantics, which stand proxy for body-grounded neural patterns. That subsymbolic models need neural grounding is also a theme of **Mortensen.**

Belew points out that because of the difference in form between the knowledge in individual connectionist networks and knowledge in science, the connectionist approach confronts a scientific barrier that the symbolic approach does not. Put differently, in its purest form the symbolic paradigm assumes that knowledge in an expert's head is a scientific theory of the domain; discovering the form of individual expert's knowledge and doing the science of the domain are the same activity. This is clearly not the case in the subsymbolic paradigm—unless we are prepared for a radically new definition of "doing the science of the domain." Belew goes on to point out that the connectionist approach places more weight on the dynamic properties of cognitive systems, rather than the static structural properties. A clear and simple example of this important point is in memory retrieval: in a traditional symbolic architecture, whether an item will be successfully retrieved depends on the static structural property of its location in memory; in a connectionist network, successful retrieval depends on whether the extended process of activation flow will settle into the desired pattern of activity.

Freeman points out that connectionist models have focussed too heavily on dynamical systems with simple static equilibria, and paid too little attention to dynamical systems with much more complex global behavior. This is undoubtedly true, but is changing, with Freeman's work on chaotic equilibria and work such as Jordan's (1986) on periodic equilibria (where "equilibria" really means "attractors"). As for the flamboyantly eliminitivist assertions ending his commentary, I think Freeman would be much harder pressed to live with the consequences of this neuro-macho talk if he were building connectionist models of language processing rather than models of olfactory pattern recognition in rabbits.

Dreyfus & Dreyfus emphasize that because the subsymbols of PTC are not necessarily context-free microfeatures, the PTC picture deviates importantly from some "language of thought" accounts. This issue is discussed in the target article in Section 7.2, but Dreyfus & Dreyfus are correct to emphasize that the distributed subconceptual representations that networks develop for themselves on their hidden units tend to be much less context-free than the example of Section 7.2 would indicate.

Lycan rejects the definition of conceptual and subconceptual levels given in the article, and so it is not surprising that he has trouble making sense of the hypotheses that crucially refer to these levels. Nonetheless his substantive comments seem to by and large support the PTC view. He points out that the "complete, formal and precise" cognitive account that PTC assumes to exist at the subconceptual level is an account at a semantically interpretable level—it's just that the interpretation is in terms of subconceptual features like "roundness preceded by frontalness and followed by backness." The **Dreyfus & Dreyfus** point just discussed entails that the typical subconceptual feature will be much more context-dependent and obscure than this, making semantic interpretation at the subconceptual level a messy business. But, as Lycan points out, the cleaner semantic interpretations residing at the conceptual level come with much more difficult computational properties. For a complex subsymbolic system, the lower level offers clean processes but messy interpretations, while the upper level offers the reverse. The clean way to go is to do semantic interpretation at the upper level and "syntax"—processing—at the lower level. The clean semantics is carried by symbols that float on the clean syntax of the subsymbols.

3. Treatment of symbolic models

It is clear that a number of the commentators sought in the target article arguments that a connectionist formalism is in principle superior to a symbolic formalism. These people were particularly disappointed with my treatment of the "symbolic paradigm" and were quick to point out that arguments against the "symbolic paradigm" as I characterized it are not arguments against symbolic computation more generally construed (eg., **Chandrasekaran, Goel, & Allemang, Lycan, Pinker & Prince, Rey, Van Gulick**)

There is a good reason that I did not try to set the discussion in terms of symbolic vs. connectionist formalisms, each broadly construed. I am convinced that such a discussion is, at least at this point, fruitless. As was spelled out in the target article in Section 2.4, symbolic computation broadly construed can be used to implement connectionist computation: that is what I do whenever I simulate a connectionist model in Lisp. In turn, connectionist computation, broadly construed, can be used to implement a Turing machine, and so all of symbolic computation.

Thus for a meaningful discussion of the relation between connectionist and symbolic models, something less than the broadest construals of the approaches must be put on the table. For each of the approaches, I identified a single, coherent approach to cognitive modeling that has a significant number of practitioners and scientific interest. I coined a term, "subsymbolic," for the connectionist approach, but did not have the corresponding foresight to coin a term for the symbolic approach, instead giving it the generic name "the symbolic paradigm." Explicit, repeated disclaimers did not suffice to convey the deliberately restricted nature of what I called "the symbolic paradigm."

It is not the role of commentators to redefine the grounds for debate, as **Pinker & Prince**, for example, explicitly attempt to do. The target article is not, and explicitly does not claim to be, an analysis of the relation between connectionist and linguistic theories. That is the ground on which Pinker & Prince and several other commentators want to take their stand; unfortunately it is not the ground on which this treatment lies.

Many commentators pointed out that conceptual vs. subconceptual levels and symbolic vs. connectionist computation are independent dimensions; that symbolic computation does not commit one to working at the conceptual level (eg., **Chandrasekaran, Goel, & Allemang, Lindsay, Lycan, Pinker & Prince, Rey, Van Gulick**) This independence is explicitly acknowledged, and even emphasized, in the target article. In (4) and (8), the independence is manifest in distinguishing the "semantic" from the "syntactic" (processing) assumptions of the two paradigms being defined. I insisted in Section 2.4 that unless the syntactic assumptions are *supplemented* (read: "independent assumption added") by semantic ones, the discussion immediately degenerates to the trivial mutual-implementability that a few paragraphs ago was cited as the reason for avoiding the most general characterization of the two approaches.

The independence of semantic levels and type of computational processes is again indicated by the two-dimensional format of Table 2. Since the syntactic and semantic assumptions are independent, there is a two-by-two space of modeling approaches, on which the "semantic axis" is the semantic level at which the model's processing is defined (conceptual or subconceptual) and on which the other "syntactic axis" is the type of computation used (symbolic or connectionist). The "symbolic paradigm" occupies the conceptual/symbolic corner, and the "subsymbolic paradigm" occupies the opposite, subconceptual/connectionist corner. The other two corners did not figure prominently in the target article. One is the conceptual/connectionist approach: local connectionism, mentioned in passing as (9). The other is the subconceptual/symbolic approach typified by much linguistics-based theory, and explicitly excluded from the scope of the target article.

The subconceptual/symbolic approach is difficult to address because it the least constrained of all. The symbols manipulated can represent arbitrarily fine-grained features, and the operations performed can be arbitrarily complex symbol manipulations. Certainly such an unconstrained approach cannot be lacking in power relative to any of the others, since in a sense all the others are special cases of it. For example, as discussed in Section 2.4, a Lisp program simulating a subsymbolic model is not a model of the symbolic paradigm, in the sense of (4), but it certainly is a model of this most general symbolic approach.

Why would a theorist willingly forgo a framework that is completely unconstrained for one that is much more constrained? Theorists do this all the time, when they are committed to a working hypothesis; they believe that the constraints willingly accepted will serve to guide them to valid accounts. The jury won't be in on the connectionist constraints for some time. But it certainly isn't valid to argue against the subsymbolic approach solely on the basis that it is more constrained. The theme "symbolic computation (most broadly construed) can do whatever connectionism can do" is a triviality. (**Nelson** *seems* to be making such a point, though I am honestly not sure.) The point is that by accepting the constraints that they do, connectionists have been led to interesting learning and processing principles that could in principle have been, but in practice were not, discovered by theorists who do not willingly accept the constraints that connectionism imposes.

While many commentators wanted to quickly dismiss the (conscious) conceptual level as irrelevant to characterizing the symbolic approach, there is a strong tradition of cognitive modeling and philosophical analysis that fits squarely within the symbolic paradigm as defined in (4). For example, models of skill acquisition (eg. Anderson 1983) in terms of internalization of taught rules, followed by rule compilation and chunking, start with taught rules that must of necessity rely on consciously accessible concepts, and then manipulate these rules in ways that never go below the conscious level towards the subconceptual. Philosophical arguments from the structure of mental states, like those championed by Fodor and Pylyshyn and presented in the commentary of **Rey**, apply at, and only at, the level of conscious thoughts. Chomsky has made it fashionable to deny the relevance of conscious access, but these arguments cannot survive without it.

4. Adequacy of connectionism in practice

There are a lot of people out there who are deeply annoyed by the outlandish claims being made in some quarters about the accomplishments and power of connectionism. This impatience is due in no small part to having listened to such claims about symbolic AI for the past 30 years. I am one of these annoyed people, and the target article contains no claims about the power of connectionism, which is, at this point, essentially completely unknown. The statements in (1) were explicitly labeled as my personal beliefs, not as claims, included only in the hope of increasing the clarity of the paper.

Just the same, a number of commentators took this opportunity to address perceived inadequacies in the power of connectionist models.

4.1. Pinker & Prince

Pinker & Prince start by refusing to respect my right to define the grounds of my analysis to exclude linguistic-based models; they then go on to accuse me of conflating a number of issues. I am perfectly aware that symbolic computation can incorporate subconceptual features, parallel processing, and, they might have gone on, soft constraints, non-monotonic logic, and fuzzy sets. The irrelevance of all this for the present treatment of the target article was just spelled out in the preceding paragraphs.

In reference to my comments comparing inference with soft and hard constraints (Section 8), **Pinker & Prince** make the elementary point that adding a new rule to a system can radically change the "ecology of the grammar," and that rule interaction creates a kind of context-dependence. My point was simply that hard constraints can be used one at time to make inferences, whereas soft constraints cannot. Given p and $p \rightarrow q$, we can conclude q with certainty, without giving any consideration to whatever other rules may be in the system. By contrast, if we know that node i is active, and that there is a positive connection from i to j, we can't conclude *anything* about the activity of node j until we know what all the other connections and activities are. This difference has important implications for performing inferences with hard and soft constraints, and is true despite the obvious fact that the *total* set of inferences using hard rules depends on the *total* set of rules.

Pinker & Prince go on to innumerate what they take to be several problems for connectionist systems. The first is that an "entity is nothing but its features"; they base this on the Rumelhart & McClelland (1986) model of past-tense acquisition. But it has been emphasized within the connectionist approach for some time (eg., Hinton 1981), that in general it is important to have "microfeatures" that serve to hold arbitrary patterns providing names for entities: there is nothing intrinsic to the connectionist approach that forbids, for example, the pattern representing a verb stem to consist in part in a pattern encoding the phonological form and in part a pattern that serves as a unique identifier for that stem (eg., to distinguish homonyms). In fact, arguments such as those in the commentary of **Dreyfus & Dreyfus** imply that such microfeatures are to be expected to be found among those invented by networks in their hidden units.

Next, **Pinker & Prince** accuse me of "bait-and-switch" because subsymbols are supposed to be more fine-grained or abstract than symbols, yet I call Wickelfeatures, which combine features of an entity with features of its context, "subsymbols." It is hard to see any duplicity or contradiction here, since in Section 7.2 I am quite explicit about the appearance of context in subsymbols. There is nothing about fine-grainedness that is inconsistent with context-sensitive subsymbols.

Pinker & Prince are quite correct that subsymbols adequate for connectionist processing are difficult to discover and not identical with the subconceptual features of symbolic theory. That is why the subconceptual level of the subsymbolic paradigm is a new level, distinct from that of fine-grained features in symbolic theory. And that is why there is so much interest in connectionist learning systems that discover their own subsymbols; the necessary technology for this was discovered *after* the development of the model they base their entire critique upon.

Pinker & Prince are concerned about connectionist networks blending competing possible outputs and thereby creating nonsense. Again, this is a real problem, and one for which solutions exist. (For example, see the discussion of phase transitions in Smolensky 1986a).

They next worry about selectively ignoring similarity. They say that because there is behavior that looks all-or-none, only mechanisms that are all-or-none can do the job. Of course, the whole point of the subsymbolic approach is to *explain* how symbolic processes, eg. all-or-none processes, can emerge from processes that are fundamentally soft and statistical. It simply does not provide any *explanation* to say that the reason there are all-or-none behaviors (sometimes) is that there are all-or-none processes.

There is no need to further pursue **Pinker & Prince**'s commentary in detail, as the loop continues: Here's something X that's easy for a symbolic model; gee, X is hard for a connectionist model; look at the Rumelhart & McClelland model's first stab at trying to do X; complain that it isn't good enough; conclude that connectionist models can't possibly do X—in fact, conclude that "connectionist models that are restricted to associations among subsymbols are demonstrably inadequate." In every case, it is true that connectionist models don't yet do X well enough, that research is underway on how to do X better, and that the state of the art is already several years beyond what Pinker & Prince critique. The conclusion that connectionist models are "demonstrably inadequate," on the basis of investigation of a single model representative of the previous generation of connectionist technology, is so grossly premature that only priests of the High Church blessed with divine revelation could claim adequate grounds for the prophesy. I wish I were so blessed, because, as I state in no uncertain terms in (1), at present it seems quite unknowable whether connectionist models can adequately solve the problems they face.

4.2. Freidin

Freidin treats us to a rerun of Chomsky's greatest hits. That old favorite, the poverty of the stimulus, is a purely intuitive argument with no formal backing in the absence of hypothesized mechanisms to test it. Fans of the argument must be delighted to see that connectionism is working its way to a position where the argument can be put to a new formal test. Freidin reminds us of the familiar point that a crucial aspect of the learnability of language is the learnability of the abstractions to which linguistic regularities are sensitive—or functionally equivalent abstractions. It will no doubt then be cause for satisfaction that a main activity of connectionist research is the study of the learnability of abstractions. The problem of distinguishing ungrammatical sentences from novel grammatical sentences is of course a special case of the problem of inductive generalization, and not at all special to language. This problem, too, figures centrally in connectionist research; every typical learning connectionist system that has ever been built has solved, with greater or lesser success, this problem. The standard learning paradigm is to choose a set of target patterns to be learned—the "grammatical sentences"—, to train the network on a subset of these patterns (no ungrammatical sentences presented!), and finally to test whether the trained network generalizes correctly to distinguish the unseen target patterns from the non-targets, ie., to distinguish "novel grammatical sentences" from "nongrammatical sentences." Success of course depends on the regularities that distinguish the "grammatical" and "ungrammatical" cases, and the representativeness of the training set.

Freidin takes the usual delight in pointing out that a connectionist model, PARSNIP, that successfully learns to model performance on grammaticality judgements, doesn't learn "grammar," for the deep reason that Chomsky has patented the term to apply to something else. What is a bit puzzling is that Freidin indulges in this delight less than a half dozen paragraphs after delivering the authorative pronouncement that when it comes to building a connectionist model of linguistic performance, "there is no reason to believe that such a model will succeed."

Before leaving **Freidin**'s commentary an obvious comment about innateness is required. A symbolic view of language acquisition currently popular in the conceptual neighborhood of Cambridge MA involves an innately-specified parametrized set of grammars together with an empirical hypothesis-testing phase of parameter adjustment. There is no a priori connectionist or even PTC view of how the learning of language is to be divided between innate and acquired knowledge. In a very literal sense, every connectionist learning network is an innately-specified parametrized set of possible knowledge configurations together with an empirical hypothesis-testing phase of parameter adjustment. The difference is that instead of a few discrete parameters imbedded in complex symbolic rules, the innate endowment is a lot of continuous parameters imbedded in simple numerical "rules." There is no *obvious* way in which the abstractions entering in the symbolic innate rules can be imbedded in the innate structure of a connectionist network, but it is far to early to tell whether there is a non-obvious way that something equivalent can—and should—be done.

4.3. Shepard

In a related vein, Shepard argues that the connectionist approach has systematically neglected an essential question: How does adaptation on an evolutionary time scale configure networks so that they are innately enabled to learn what they must learn? I believe Shepard is correct both in characterizing this as lack and in emphasizing its importance. The neglect is probably a result of the lack of any technical handle on the problem; this is obviously a fertile ground for someone with the right set of tools. (**Chandrasekaran, Goel, & Allemang** raise the same issue of grappling with the prior commitments imbedded in network architectures.)

4.4. Rey

Rey wonders if patterns of activity can be used to create mental states with the properties he demands in his (1)-(4). I believe that the approach laid out in Smolensky (1987), which constructs and analyzes fully distributed structured representations composed from subpatterns in appropriate ways, can get close enough to do the necessary work. (Indeed it was exactly considerations such as Rey's (1)-(4), impressed upon me by Rey, Fodor, and Pylyshyn, that motivated this research.)

4.5. Lehnert

Lehnert's commentary makes the following points:

a. The symbolic/subsymbolic debate is tedious.
b. Psychologists are attracted to connectionism because of theorem envy.
c. Connectionism is methodology driven research and that's dangerous.
d. The methodology driving Smolensky is physics.
e. Smolensky wants to ignore representational issues.

The fact that Lehnert bothered to comment on the target article tends to undercut an otherwise dramatic opening with point a. A good ad hominem attack in point b., however, picks up the credibility. The presupposition of this attack is, unfortunately, patently false: the symbolic approach offers many more theorems, both in absolute numbers and in current rate of production, than the connectionist approach (see any issue of the journal *Artificial Intelligence*); it just so happens that the school of symbolic AI to which Lehnert belongs prefers to regard such theorems as irrelevant.

Point c. is not argued, simply asserted. It does not seem a correct accusation, but for the sake of argument, I will accept it nonetheless. With a formalism as undeveloped as connectionism, anyone who thinks the approach will get very far without considerable attention to methodological problems is, I think, quite naive about the maturity required of a formalism to be adequate for cognitive modeling. Symbolic computation was developed through decades of methodologically driven research, and researchers who now want to apply it can afford the luxury of focussing exclusively on problem domains. The connectionist community as a whole cannot now afford that luxury. Some of us have to worry about the methodologically driven problems, and we each apply the tools that we think we can get the most mileage from. Mine happen to be tools from physics and I'm sorry if **Lehnert** finds that offensive.

Where **Lehnert** gets the idea that I believe representational issues are not central to connectionism is beyond me. I am unaware of any paper that devotes more attention than the target article to foundational questions of connectionist representation. As for technical attention to connectionist representation: "For most ... aspects of connectionist modeling, there exists considerable formal literature analyzing the problem and offering solutions. There is one glaring exception: the representation component. This is a crucial component, for a poor representation will often doom the model to failure, and an excessively generous representation may essentially solve the problem in advance. Representation is particularly critical to understanding the relation between connectionist and symbolic computation, for the representation often embodies most of the relation between a symbolically characterized problem (eg. a linguistic task) and a connectionist solution." This is quote from

Smolensky (1987), "On Variable Binding and the Representation of Symbolic Structures in Connectionist Systems," which sets out a general mathematical framework for analyzing the problem of connectionist representation, and defines and analyzes a general technique for representing structured data, such as stacks and trees. That this paper's results are presented as theorems would surely offend Lehnert's sensibilities, but the work should leave little doubt about the importance I attach to issues of connectionist representation.

4.6. Hunter

Hunter first claims my definition of connectionism is too broad; he next says that my claims "are perhaps best taken to refer" to a very narrowly (and self-contraditorily)[6] defined set of networks; Hunter then proceeds in the bulk of the commentary to argue that these networks are much too narrow to constitute a general framework for cognition. My claims are, in fact, "best taken to refer" to exactly the systems I defined them to refer to, and not to the small set Hunter appears to be familiar with. Whatever weaknesses the target article may have, extreme narrowness of the framework is not among them. Since Hunter has chosen to comment on a target article of his own imagining, I'll leave the response to his imagination as well.

4.7. McCarthy

McCarthy's comments about unary fixation and lack of "elaboration tolerance" seem to be on target. At this point, connectionist models tend to be developed with the absolute bare minimum (or slightly less) of representational power for the target task. If the task is beefed up even a little bit, the model has to be scrapped. This is a sign of the immaturity of connectionist representations; it is hard enough to get one that is barely adequate—doing more is not usually entertained.

It would be a mistake to leave the impression that connectionist models cannot represent higher than unary relations. One technique involves binding arguments of a relation to the slots in the relation. Research on this problem includes Hinton's (1981) work on semantic networks, McClelland & Kawamoto's (1986) work on semantic role assignment, Derthick's (1986, 1987) connectionist implementation of a micro-KL-ONE, Touretzky & Hinton's (1985) connectionist implementation of a simple production system, and my work (Smolensky 1987) on representation of symbolic structures. In much of this work, the key is to use micro-features that denote the conjunction of a micro-feature of the slot and a microfeature of the argument filling that slot: greater-than-unary relations are achieved by using greater-than-unary microfeatures. For greater-than-unary analysis of rooms, the network would need to be trained not on patterns describing single rooms in isolation, but patterns describing configurations of rooms, with the necessary interrelations included in the descriptions.

As for the proper connectionist treatment of an English speaker pronouncing Chinese names, the analysis implicitly proposed in the target article is the following. Rules about how to pronounce Chinese *Q* and *X* are entered into the (connectionist-implemented) conscious rule interpreter as S-knowledge (Section 6.3); resident in the intuitive processor are the NETtalk-like connections constituting P-knowledge of English pronounciation. When the English speaker is reading English text, the computation is done in parallel in the intuitive processor. When a Chinese name comes along, the intuitive processor fails to settle quickly on a pronounciation because of the non-English-like letter sequences; in particular, sequences starting with *Q* or *X* are likely to be quite unpronounceable, as far as this P-knowledge is concerned. Because the intuitive processor has failed to quickly settle on a pronunciation, the rule interpreter has a chance to carry out the rather slow process of firing its rules for pronouncing *Q* and *X*. With some practice, the intuitive processor starts to tune its weights to handle these new cases: the S-knowledge is slowly and indirectly "compiled" into P-knowledge.

This account makes a number of predictions about performance: pronunciation of *Q*- and *X*-names will (initially) be accompanied by conscious awareness of use of the rules, can be interfered with by other conscious processes, and will have an identifiably different time course; many more mistakes would be made if, instead of *Q*

[6] Simulated annealing is not really a training technique, and the method he presumably means, Boltzmann learning, can't really be used with feedforward networks.

and X, letters were used that are pronounceable in English in a larger variety of contexts (eg., if T and K were pronounced D and G), and so forth.

Note that the proper treatment of this task does *not* involve instantaneously adjusting the connections in the NETtalk-like intuitive processor to incorporate the pronounciation of Q and X; these connections are established only slowly through practice. But at instruction time, it *is* necessary to instantly change many connections in the conscious rule interpreter in order to store the new rules. How this might be done is the subject of current research, but note that it is only the special-purpose conscious rule interpreter, built on the language processor, that needs to perform one-trial learning; specialized intuitive modules do not need this capability. The basic idea for how the language processor can handle one-trial learning is this. The capability to understand a language requires a network that has many "virtual harmony maxima" at states corresponding to the well-formed sentences and their meanings; when a well-formed sentence is heard, even once, the prior tuning of the network to the language enables the network to turn the "virtual harmony maximum" corresponding to that sentence into a real harmony maximum: a stored memory. Whether this proposal can actually be carried out is unknown at this time.

References*

Derthick, M. (1986) A connectionist knowledge representation system. Thesis proposal, Computer Science Department, Carnegie-Mellon University.

Derthick, M. (1987) A connectionist architecture for representing and reasoning about structured knowledge. *Proceedings of the Ninth Meeting of the Cognitive Science Society.*

Hinton, G.E. (1981) Implementing semantic networks in parallel hardware. In: *Parallel models of associative memory,* G.E. Hinton & J.A. Anderson (Eds.) Erlbaum.

Hofstadter, D.R. (1983) The architecture of Jumbo. *Proceedings of the International Machine Learning Workshop.*

McClelland, J.L. & Kawamoto, A.H. (1986) Mechanisms of sentence processing: Assigning roles to constituents. In: *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 2: Psychological and biological models,* J. L. McClelland, D. E. Rumelhart, & the PDP Research Group. MIT Press/Bradford Books.

Pinker, S. & Prince, A. (1987) On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. Occasional paper #33, Center for Cognitive Science, MIT.

Smolensky, P. (1987) On variable binding and the representation of symbolic structures in connectionist systems. Technical Report CU–CS–355–87, Department of Computer Science, University of Colorado at Boulder. Revised version to appear in *Artificial Intelligence.*

*References already included with the target article are not replicated here.

	Position				
	Eliminativist		Limitivist	Revisionist	Implementationalist
	Neural	Connectionist	(PTC)		
Conceptual level laws/ Symbolic processes	folklore	folklore	approximately correct	exactly correct, after revision	exactly correct
Subconceptual level laws/ Connectionist processes	nonexistent	exactly correct for entire cognitive system	exactly correct for entire cognitive system	exactly correct (for connectionist part of cognitive system)	exactly correct (but irrelevant for cognitive architecture)

Table A: A spectrum of positions on connectionism's relation to the symbolic approach.

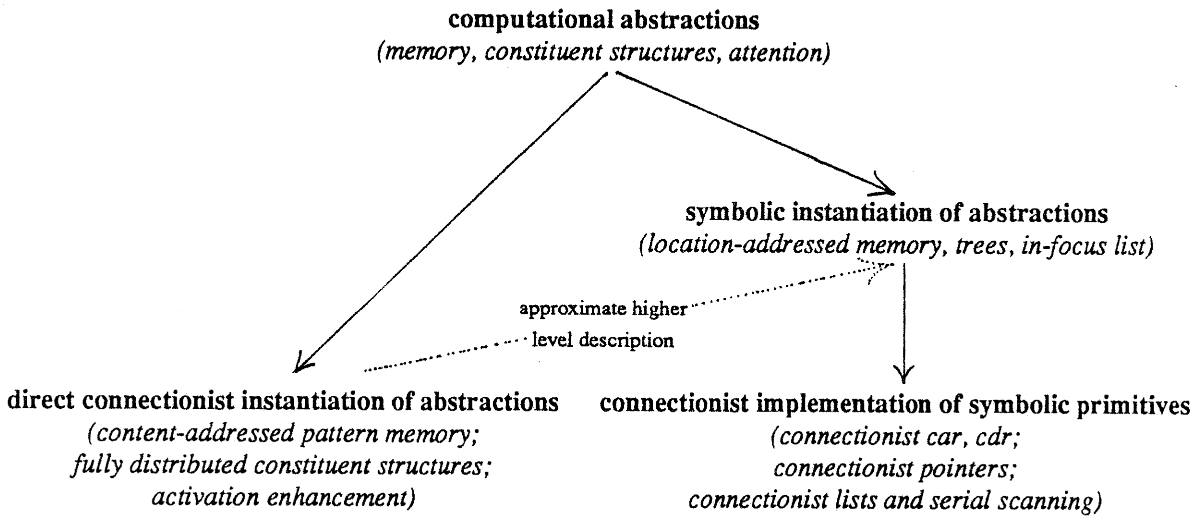
computational abstractions
(memory, constituent structures, attention)
symbolic instantiation of abstractions
(location-addressed memory, trees, in-focus list)
approximate higher
level description
direct connectionist instantiation of abstractions
(content-addressed pattern memory;
fully distributed constituent structures;
activation enhancement)
connectionist implementation of symbolic primitives
(connectionist car, cdr;
connectionist pointers;
connectionist lists and serial scanning)

Figure A: The methodological implications of limitivist (left branch) vs. implementationalist (right branch) views of connectionism.