BEHAVIORAL AND BRAIN SCIENCES (1988) 11, 1–74
Printed in the United States of America

On the proper treatment of connectionism

Paul Smolensky
Department of Computer Science and Institute of Cognitive Science, University of Colorado, Boulder, Colo. 80309-0430

Abstract: A set of hypotheses is formulated for a connectionist approach to cognitive modeling. These hypotheses are shown to be incompatible with the hypotheses underlying traditional cognitive models. The connectionist models considered are massively parallel numerical computational systems that are a kind of continuous dynamical system. The numerical variables in the system correspond semantically to fine-grained features below the level of the concepts consciously used to describe the task domain. The level of analysis is intermediate between those of symbolic cognitive models and neural models. The explanations of behavior provided are like those traditional in the physical sciences, unlike the explanations provided by symbolic models.

Higher-level analyses of these connectionist models reveal subtle relations to symbolic models. Parallel connectionist memory and linguistic processes are hypothesized to give rise to processes that are describable at a higher level as sequential rule application. At the lower level, computation has the character of massively parallel satisfaction of soft numerical constraints; at the higher level, this can lead to competence characterizable by hard rules. Performance will typically deviate from this competence since behavior is achieved not by interpreting hard rules but by satisfying soft constraints. The result is a picture in which traditional and connectionist theoretical constructs collaborate intimately to provide an understanding of cognition.

Keywords: cognition; computation; connectionism; dynamical systems; networks; neural models; parallel distributed processing; symbolic models

1. Introduction

In the past half-decade the connectionist approach to cognitive modeling has grown from an obscure cult claiming a few true believers to a movement so vigorous that recent meetings of the Cognitive Science Society have begun to look like connectionist pep rallies. With the rise of the connectionist movement come a number of fundamental questions which are the subject of this target article. I begin with a brief description of connectionist models.

1.1. Connectionist models. Connectionist models are large networks of simple parallel computing elements, each of which carries a numerical *activation value* which it computes from the values of neighboring elements in the network, using some simple numerical formula. The network elements, or *units*, influence each other's values through connections that carry a numerical strength, or *weight*. The influence of unit i on unit j is the activation value of unit i times the strength of the connection from i to j. Thus, if a unit has a positive activation value, its influence on a neighbor's value is positive if its weight to that neighbor is positive, and negative if the weight is negative. In an obvious neural allusion, connections carrying positive weights are called *excitatory* and those carrying negative weights are *inhibitory*.

In a typical connectionist model, input to the system is provided by imposing activation values on the *input units* of the network; these numerical values represent some encoding, or *representation*, of the input. The activation on the input units propagates along the connections until some set of activation values emerges on the *output units*; these activation values encode the output the system has computed from the input. In between the input and output units there may be other units, often called *hidden units*, that participate in representing neither the input nor the output.

The computation performed by the network in transforming the input pattern of activity to the output pattern depends on the set of connection strengths; these weights are usually regarded as encoding the system's knowledge. In this sense, the connection strengths play the role of the program in a conventional computer. Much of the allure of the connectionist approach is that many connectionist networks *program themselves*, that is, they have autonomous procedures for tuning their weights to eventually perform some specific computation. Such learning procedures often depend on training in which the network is presented with sample input/output pairs from the function it is supposed to compute. In learning networks with hidden units, the network itself "decides" what computations the hidden units will perform; because these units represent neither inputs nor outputs, they are never "told" what their values should be, even during training.

In recent years connectionist models have been developed for many tasks, encompassing the areas of vision, language processing, inference, and motor control. Numerous examples can be found in recent proceedings of the meetings of the Cognitive Science Society; *Cognitive Science* (1985); Feldman et al. (1985); Hinton and Anderson (1981); McClelland, Rumelhart, and the PDP Re-

© 1988 Cambridge University Press 0140-525X/88 $5.00+.00

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

search Group (1986); Rumelhart, McClelland, and the PDP Research Group (1986). [See also Ballard "Cortical Connections and Parallel Processing" *BBS* 9(1) 1986.]

1.2. Goal of this target article. Given the rapid development in recent years of the connectionist approach to cognitive modeling, it is not yet an appropriate time for definitive assessments of the power and validity of the approach. The time seems right, however, for an attempt to articulate the goals of the approach, the fundamental hypotheses it is testing, and the relations presumed to link it with the other theoretical frameworks of cognitive science. A coherent and plausible articulation of these fundamentals is the goal of this target article. Such an articulation is a nontrivial task, because the term "connectionist" encompasses a number of rather disparate theoretical frameworks, all of them quite undeveloped. The connectionist framework I will articulate departs sufficiently radically from traditional approaches in that its relations to other parts of cognitive science are not simple.

For the moment, let me call the formulation of the connectionist approach that I will offer *PTC*. I will not argue the scientific merit of PTC; that some version of connectionism along the lines of PTC constitutes a "proper description of processing" is argued elsewhere (e.g., in Rumelhart, McClelland & the PDP Research Group 1986; McClelland, Rumelhart & the PDP Research Group 1986). Leaving aside the scientific merit of connectionist models, I want to argue here that PTC offers a "Proper Treatment of Connectionism": a coherent formulation of the connectionist approach that puts it in contact with other theory in cognitive science in a particularly constructive way. PTC is intended as a formulation of connectionism that is at once strong enough to constitute a major cognitive hypothesis, comprehensive enough to face a number of difficult challenges, and sound enough to resist a number of objections in principle. If PTC succeeds in these goals, it will facilitate the real business at hand: Assessing the scientific adequacy of the connectionist approach, that is, determining whether the approach offers computational power adequate for human cognitive competence and appropriate computational mechanisms to accurately model human cognitive performance.

PTC is a response to a number of positions that are being adopted concerning connectionism – pro, con, and blandly ecumenical. These positions, which are frequently expressed orally but rarely set down in print, represent, I believe, failures of supporters and critics of the traditional approach truly to come to grips with each other's views. Advocates of the traditional approach to cognitive modeling and AI (artificial intelligence) are often willing to grant that connectionist systems are useful, perhaps even important, for modeling lower-level processes (e.g., early vision), or for fast and fault-tolerant implementation of conventional AI programs, or for understanding how the brain might happen to implement LISP. These ecumenical positions, I believe, fail to acknowledge the true challenge that connectionists are posing to the received view of cognition; PTC is an explicit formulation of this challenge.

Other supporters of the traditional approach find the connectionist approach to be fatally flawed because it cannot offer anything new (since Universal Turing machines are, after all, "universal"), or because it cannot offer the kinds of explanations that cognitive science requires. Some dismiss connectionist models on the grounds that they are too neurally unfaithful. PTC has been designed to withstand these attacks.

On the opposite side, most existing connectionist models fail to come to grips with the traditional approach – partly through a neglect intended as benign. It is easy to read into the connectionist literature the claim that there is no role in cognitive science for traditional theoretical constructs such as rules, sequential processing, logic, rationality, and conceptual schemata or frames. PTC undertakes to assign these constructs their proper role in a connectionist paradigm for cognitive modeling. PTC also addresses certain foundational issues concerning mental states.

I see no way of achieving the goals of PTC without adopting certain positions that will be regarded by a number of connectionists as premature or mistaken. These are inevitable consequences of the fact that the connectionist approach is still quite underdeveloped, and that the term "connectionist" has come to label a number of approaches that embody significantly conflicting assumptions. PTC is *not* intended to represent a consensus view of what the connectionist approach is or should be.

It will perhaps enhance the clarity of the article if I attempt at the outset to make my position clear on the present value of connectionist models and their future potential. This article is not intended as a defense of all these views, though I will argue for a number of them, and the remainder have undoubtedly influenced the presentation. On the one hand, I believe that:

(1) a. It is far from clear whether connectionist models have adequate computational power to perform high-level cognitive tasks: There are serious obstacles that must be overcome before connectionist computation can offer modelers power comparable to that of symbolic computation.

b. It is far from clear that connectionist models offer a sound basis for modeling human cognitive performance: The connectionist approach is quite difficult to put into detailed contact with empirical methodologies.

c. It is far from clear that connectionist models can contribute to the study of human competence: Connectionist models are quite difficult to analyze for the kind of high-level properties required to inform the study of human competence.

d. It is far from clear that connectionist models, in something like their present forms, can offer a sound basis for modeling neural computation: As will be explicitly addressed in Section 4, there are many serious gaps between connectionist models and current views of important neural properties.

e. Even under the most successful scenario for connectionist cognitive science, many of the currently practiced research strategies in cognitive science would remain viable and productive.

On the other hand, I believe that:

(1) f. It is very likely that the connectionist approach will contribute significant, long-lasting ideas to the rather impoverished theoretical repertoire of cognitive science.

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

g. It is very likely that connectionist models will turn out to offer contributions to the modeling of human cognitive performance on higher-level tasks that are at least as significant as those offered by traditional, symbolic, models.

h. It is likely that the view of the competence/performance distinction that arises from the connectionist approach will successfully heal a deep and ancient rift in the science and philosophy of mind.

i. It is likely that connectionist models will offer the most significant progress of the past several millenia on the mind/body problem.

j. It is very likely that, given the impoverished theoretical repertoire of computational neuroscience, connectionist models will serve as an excellent stimulus to the development of models of neural computation that are significantly better than both current connectionist models and current neural models.

k. There is a reasonable chance that connectionist models will lead to the development of new somewhat-general-purpose self-programming, massively parallel analog computers, and a new theory of analog parallel computation: They may possibly even challenge the strong construal of Church's Thesis as the claim that the class of well-defined computations is exhausted by those of Turing machines.

1.3. Levels of analysis. Most of the foundational issues surrounding the connectionist approach turn, in one way or another, on the level of analysis adopted. The terminology, graphics, and discussion found in most connectionist papers strongly suggest that connectionist modeling operates at the neural level. I will argue, however, that it is better *not* to construe the principles of cognition being explored in the connectionist approach as the principles of the neural level. Specification of the level of cognitive analysis adopted by PTC is a subtle matter which consumes much of this article. To be sure, the level of analysis adopted by PTC is lower than that of the traditional, symbolic paradigm; but, at least for the present, the level of PTC is more explicitly related to the level of the symbolic paradigm than it is to the neural level. For this reason I will call the paradigm for cognitive modeling proposed by PTC the *subsymbolic paradigm.*

A few comments on terminology. I will refer to the traditional approach to cognitive modeling as the *symbolic paradigm.* Note that I will always use the term "symbolic paradigm" to refer to the traditional approach to cognitive *modeling:* the development of AI-like computer programs to serve as models of psychological performance. The symbolic paradigm in cognitive modeling has been articulated and defended by Newell and Simon (1976; Newell 1980), as well as by Fodor (1975; 1987), Pylyshyn (1984), and others. The fundamental hypotheses of this paradigm embrace most of mainstream AI, in addition to AI-based systems that are explicitly offered as models of human performance. The term "symbolic paradigm" is explicitly *not* intended to encompass competence theories such as the formal theory of grammar; such competence theories bear deep relations to the symbolic paradigm but they are not a focus of attention in this paper. In particular, much of the work in formal linguistics differs from the symbolic paradigm in cognitive modeling in many of the same ways as the connectionist approach I will consider; on a number of the dimensions I will use to divide the symbolic and subsymbolic paradigms, much linguistics research falls on the subsymbolic side.

I have found it necessary to deal only with a subset of the symbolic and connectionist approaches in order to get beyond superficial, syntactic issues. On the symbolic side, I am limiting consideration to the Newell/ Simon-/Fodor/Pylyshyn view of cognition, and excluding, for example, the view adopted by much of linguistics; on the connectionist side, I will consider only a particular view, the "subsymbolic paradigm," and exclude a number of competing connectionist perspectives. The only alternative I see at this point is to characterize the symbolic and connectionist perspectives so diffusely that substantive analysis becomes impossible.

In calling the traditional approach to cognitive modeling the "symbolic paradigm," I intend to emphasize that in this approach, cognitive descriptions are built of entities that are symbols both in the semantic sense of referring to external objects and in the syntactic sense of being operated upon by symbol manipulation. These manipulations model fundamental psychological processes in this approach to cognitive modeling.

The name "subsymbolic paradigm" is intended to suggest cognitive descriptions built up of entities that correspond to *constituents* of the symbols used in the symbolic paradigm; these fine-grained constituents could be called *subsymbols,* and they are the activities of individual processing units in connectionist networks. Entities that are typically represented in the symbolic paradigm by symbols are typically represented in the subsymbolic paradigm by a large number of subsymbols. Along with this semantic distinction comes a syntactic distinction. Subsymbols are not operated upon by symbol manipulation: They participate in numerical – not symbolic – computation. Operations in the symbolic paradigm that consist of a single discrete operation (e.g., a memory fetch) are often achieved in the subsymbolic paradigm as the result of a large number of much finer-grained (numerical) operations.

Since the level of cognitive analysis adopted by the subsymbolic paradigm for formulating connectionist models is lower than the level traditionally adopted by the symbolic paradigm, for the purposes of relating these two paradigms, it is often important to analyze connectionist models at a higher level; to amalgamate, so to speak, the subsymbols into symbols. Although the symbolic and subsymbolic paradigms each have their preferred level of analysis, the cognitive models they offer can be described at multiple levels. It is therefore useful to have distinct names for the levels: I will call the preferred level of the symbolic paradigm the *conceptual level* and that of the subsymbolic paradigm the *subconceptual level.* These names are not ideal, but will be further motivated in the course of characterizing the levels. A primary goal of this article is to articulate a coherent set of hypotheses about the subconceptual level: the kind of cognitive descriptions that are used, the computational principles that apply, and the relations between the subconceptual and both the symbolic and neural levels.

The choice of level greatly constrains the appropriate formalism for analysis. Probably the most striking feature

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

of the connectionist approach is the change in formalism relative to the symbolic paradigm. Since the birth of cognitive science, *language* has provided the dominant theoretical model. Formal cognitive models have taken their structure from the syntax of formal languages, and their content from the semantics of natural language. The mind has been taken to be a machine for formal symbol manipulation, and the symbols manipulated have assumed essentially the same semantics as words of English.

The subsymbolic paradigm challenges both the syntactic and semantic role of language in formal cognitive models. Section 2 formulates this challenge. Alternative fillers are described for the roles language has traditionally played in cognitive science, and the new role left to language is delimited. The fundamental hypotheses defining the subsymbolic paradigm are formulated, and the challenge that nothing new is being offered is considered. Section 4 considers the relation between the subsymbolic paradigm and neuroscience; the challenge that connectionist models are too neurally unfaithful is addressed. Section 5 presents the relations between analyses of cognition at the neural, subconceptual, and conceptual levels. It also previews the remainder of the article, which deals with the relations between the subconceptual and conceptual levels; the types of explanations of behavior provided by the symbolic and subsymbolic paradigms are then discussed. Section 6 faces the challenge of accounting for conscious, rule-guided behavior within the subsymbolic paradigm. Section 7 addresses the challenge of distinguishing cognitive from noncognitive systems at the subconceptual level. Various properties of subsymbolic mental states, and the issue of rationality, are considered. Section 8 elaborates briefly on the computational principles that apply at the subconceptual level. Section 9 discusses how higher, conceptual-level descriptions of subsymbolic models approximate symbolic models (under their conceptual-level descriptions).

In this target article I have tried to typographically isolate concise formulations of the main points. Most of these numbered points serve to characterize the subsymbolic paradigm, but a few define alternative points of view; to avoid confusion, the latter have been explicitly tagged by the phrase, *To be rejected.*

2. Formalization of knowledge

2.1. Cultural knowledge and conscious rule interpretation. What is an appropriate formalization of the knowledge that cognitive agents possess and the means by which they use that knowledge to perform cognitive tasks? As a starting point, we can look to those knowledge formalizations that predate cognitive science. The most formalized knowledge is found in sciences like physics that rest on mathematical principles. Domain knowledge is formalized in linguistic structures such as "energy is conserved" (or an appropriate encryption), and logic formalizes the use of that knowledge to draw conclusions. Knowledge consists of axioms, and drawing conclusions consists of proving theorems.

This method of formulating knowledge and drawing conclusions has extremely valuable properties:

(2) a. *Public access*: The knowledge is accessible to many people.
b. *Reliability*: Different people (or the same person at different times) can reliably check whether conclusions have been validly reached.
c. *Formality, bootstrapping, universality*: The inferential operations require very little experience with the domain to which the symbols refer.

These three properties are important for science because it is a cultural activity. It is of limited social value to have knowledge that resides purely in one individual (2a). It is of questionable social value to have knowledge formulated in such a way that different users draw different conclusions (e.g., can't agree that an experiment falsifies a theory) (2b). For cultural propagation of knowledge, it is helpful if novices with little or no experience with a task can be given a means for performing that task, and thereby a means for acquiring experience (2c).

There are cultural activities other than science that have similar requirements. The laws of a nation and the rules of an organization are also linguistically formalized procedures for effecting action which different people can carry out with reasonable reliability. In all these cases, the goal is to create an abstract decision system that resides outside any single person.

Thus, at the cultural level, the goal is to express knowledge in a form that can be executed reliably by different people, even inexperienced ones. We can view the top-level conscious processor of individual people as a *virtual machine* – the *conscious rule interpreter* – and we can view cultural knowledge as a program that runs on that machine. Linguistic formulations of knowledge are perfect for this purpose. The procedures that different people can reliably execute are explicit, step-by-step linguistic instructions. This is what has been formalized in the theory of *effective procedures* (Turing 1936). Thanks to property (2c), the top-level conscious human processor can be idealized as universal: capable of executing any effective procedure. The theory of effective procedures – the classical theory of computation (Hopcroft & Ullman, 1979) – is physically manifest in the von Neumann (serial) computer. One can say that the von Neumann computer is a machine for automatically following the kinds of explicit instructions that people can fairly reliably follow – but much faster and with perfect reliability.

Thus we can understand why the production system of computation theory, or more generally the von Neumann computer, has provided a successful model of how people execute instructions (e.g., models of novice physics problem solving such as that of Larkin et al. 1980). In short, when people (e.g., novices) consciously and sequentially follow rules (such as those they have been taught), their cognitive processing is naturally modeled as the sequential interpretation[1] of a linguistically formalized procedure. The rules being followed are expressed in terms of the consciously accessible concepts with which the task domain is conceptualized. In this sense, the rules are formulated at the conceptual level of analysis.

To sum up:

(3) a. Rules formulated in natural language can provide an effective formalization of cultural knowledge.

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

b. Conscious rule application can be modeled as the sequential interpretation of such rules by a virtual machine called the conscious rule interpreter.
c. These rules are formulated in terms of the concepts consciously used to describe the task domain – they are formulated at the conceptual level.

2.2. Individual knowledge, skill, and intuition in the symbolic paradigm. The constraints on cultural knowledge formalization are not the same as those on individual knowledge formalization. The intuitive knowledge in a physics expert or a native speaker may demand, for a truly accurate description, a formalism that is not a good one for cultural purposes. After all, the individual knowledge in an expert's head does not possess the properties (2) of cultural knowledge: It is not publically accessible or completely reliable, and it is completely dependent on ample experience. Individual knowledge is a program that runs on a virtual machine that need not be the same as the top-level conscious processor that runs the cultural knowledge. By definition, conclusions reached by intuition do not come from conscious application of rules, and intuitive processing need not have the same character as conscious rule application.

What kinds of programs are responsible for behavior that is not conscious rule application? I will refer to the virtual machine that runs these programs as the *intuitive processor*. It is presumably responsible for all of animal behavior and a huge portion of human behavior: Perception, practiced motor behavior, fluent linguistic behavior, intuition in problem solving and game playing – in short, practically all skilled performance. The transference of responsibility from the conscious rule interpreter to the intuitive processor during the acquisition of skill is one of the most striking and well-studied phenomena in cognitive science (Anderson 1981). An analysis of the formalization of knowledge must consider both the knowledge involved in novices' conscious application of rules and the knowledge resident in experts' intuition, as well as their relationship.

An appealing possibility is this:

(4) a. The programs running on the intuitive processor consist of linguistically formalized rules that are sequentially interpreted. (*To be rejected.*)

This has traditionally been the assumption of cognitive science. Native speakers are unconsciously interpreting rules, as are physics experts when they are intuiting answers to problems. Artificial intelligence systems for natural language processing and problem solving are programs written in a formal language for the symbolic description of procedures for manipulating symbols.

To the syntactic hypothesis (4a) a semantic one corresponds:

(4) b. The programs running on the intuitive processor are composed of elements, that is, symbols, referring to essentially the same concepts as the ones used to consciously conceptualize the task domain. (*To be rejected.*)

This applies to production system models in which the productions representing expert knowledge are compiled versions of those of the novice (Anderson 1983; Lewis 1978) and to the bulk of AI programs.

Hypotheses (4a) and (4b) together comprise:

(4) The unconscious rule interpretation hypothesis: (*To be rejected.*)
The programs running on the intuitive processor have a syntax and semantics comparable to those running on the conscious rule interpreter.

This hypothesis has provided the foundation for the symbolic paradigm for cognitive modeling. Cognitive models of both conscious rule application and intuitive processing have been programs constructed of entities which are *symbols* both in the syntactic sense of being operated on by symbol manipulation and in the semantic sense of (4b). Because these symbols have the conceptual semantics of (4b), I am calling the level of analysis at which these programs provide cognitive models the *conceptual level.*

2.3. The subsymbolic paradigm and intuition. The hypothesis of unconscious rule interpretation (4) is an attractive possibility which a connectionist approach to cognitive modeling rejects. Since my purpose here is to formulate rather than argue the scientific merits of a connectionist approach, I will not argue against (4) here. I will point out only that in general, connectionists do not casually reject (4). Several of today's leading connectionist researchers were intimately involved with serious and longstanding attempts to make (4) serve the needs of cognitive science.[2] Connectionists tend to reject (4) because they find the consequences that have actually resulted from its acceptance to be quite unsatisfactory, for a number of quite independent reasons, including:

(5) a. Actual AI systems built on hypothesis (4) seem too brittle, too inflexible, to model true human expertise.
b. The process of articulating expert knowledge in rules seems impractical for many important domains (e.g., common sense).
c. Hypothesis (4) has contributed essentially no insight into how knowledge is represented in the brain.

What motivates the pursuit of connectionist alternatives to (4) is a hunch that such alternatives will better serve the goals of cognitive science. Substantial empirical assessment of this hunch is probably at least a decade away. One possible alternative to (4a) is:

(6) The neural architecture hypothesis: (*To be rejected.*)
The intuitive processor for a particular task uses the same architecture that the brain uses for that task.

Whatever appeal this hypothesis might have, it seems incapable in practice of supporting the needs of the vast majority of cognitive models. We simply do not know what architecture the brain uses for performing most cognitive tasks. There may be some exceptions (such as visual and spatial tasks), but for problem solving, language, and many others (6) simply cannot do the necessary work at the present time.

These points and others relating to the neural level will be considered in more detail in Section 4. For now the point is simply that characterizing the level of analysis of connectionist modeling is not a matter of simply identifying it with the neural level. While the level of analysis adopted by most connectionist cognitive models is not the conceptual one, it is also not the neural level. [See also

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

Anderson: "Methodologies for Studying Human Knowledge" *BBS* 10(3) 1987.]

The goal now is to formulate a connectionist alternative to (4) that, unlike (6), provides a viable basis for cognitive modeling. A first, crude approximation to this hypothesis is:

(7) The intuitive processor has a certain kind of connectionist architecture (which abstractly models a few of the most general features of neural networks). (To be elaborated.)

Postponing consideration of the neural issues to Section 4, we now consider the relevant kind of connectionist architecture.

The view of the connectionist architecture I will adopt is the following (for further treatment of this viewpoint, see Smolensky 1986b). The numerical activity values of all the processors in the network form a large *state vector*. The interactions of the processors, the equations governing how the activity vector changes over time as processors respond to one another's values, is an *activation evolution equation*. This evolution equation governing the mutual interactions of the processors involves the connection weights: numerical parameters which determine the direction and magnitude of the influence of one activation value on another. The activation equation is a differential equation (usually approximated by the finite difference equation that arises from discrete time slices; the issue of discrete approximation is taken up in Section 8.1). In learning systems, the connection weights change during training according to the learning rule, which is another differential equation: the *connection evolution equation*.

Knowledge in a connectionist system lies in its connection strengths. Thus, for the first part of our elaboration on (7) we have the following alternative to (4a):

(8) a. The connectionist dynamical system hypothesis:
The state of the intuitive processor at any moment is precisely defined by a vector of numerical values (one for each unit). The dynamics of the intuitive processor are governed by a differential equation. The numerical parameters in this equation constitute the processor's program or knowledge. In learning systems, these parameters change according to another differential equation.

This hypothesis states that the intuitive processor is a certain kind of *dynamical system*: Like the dynamical systems traditionally studied in physics, the state of the system is a numerical vector evolving in time according to differential evolution equations. The special properties that distinguish this kind of dynamical system – a *connectionist dynamical system* – are only vaguely described in (8a). A much more precise specification is needed. It is premature at this point to commit oneself to such a specification, but one large class of subsymbolic models is that of quasilinear dynamical systems, explicitly discussed in Smolensky (1986b) and Rumelhart, Hinton, and Williams (1986). Each unit in a quasilinear system computes its value by first calculating the weighted sum of its inputs from other units and then transforming this sum with a nonlinear function. An important goal of the subsymbolic paradigm is to characterize the computational properties of various kinds of connectionist dynamical systems (such as quasilinear systems) and thereby determine which kinds provide appropriate models of various types of cognitive processes.

The connectionist dynamical system hypothesis (8a) provides a connectionist alternative to the syntactic hypothesis (4a) of the symbolic paradigm. We now need a semantic hypothesis compatible with (8a) to replace (4b). The question is: What does a unit's value *mean*? The most straightforward possibility is that the semantics of each unit is comparable to that of a word in natural language; each unit represents such a concept, and the connection strengths between units reflect the degree of association between the concepts.

(9) The conceptual unit hypothesis: (*To be rejected.*)
Individual intuitive processor elements – individual units – have essentially the same semantics as the conscious rule interpreter's elements, namely, words of natural language.

But (8a) and (9) make an infertile couple. Activation of concepts spreading along degree of association links may be adequate for modeling simple aspects of cognition – such as relative times for naming words or the relative probabilities of perceiving letters in various contexts – but it cannot be adequate for complex tasks such as question answering or grammaticality judgments. The relevant structures cannot even be feasibly represented in such a network, let alone effectively processed.

Great computational power must be present in the intuitive processor to deal with the many cognitive processes that are extremely complex when described at the conceptual level. The symbolic paradigm, based on hypothesis (4), gets its power by allowing highly complex, essentially arbitrary, operations on symbols with conceptual-level semantics: simple semantics, complex operations. If the operations are required to be as simple as those allowed by hypothesis (8a), we cannot get away with a semantics as simple as that of (9).[3] A semantics compatible with (8a) must be more complicated:

(8) b. The subconceptual unit hypothesis:
The entities in the intuitive processor with the semantics of conscious concepts of the task domain are complex patterns of activity over many units. Each unit participates in many such patterns.

(See several of the papers in Hinton & Anderson 1981; Hinton, McClelland & Rumelhart 1986; the neural counterpart is associated with Hebb 1949; Lashley 1950, about which see Feldman 1986.) The interactions between *individual units* are simple, but these units do not have conceptual semantics: they are *subconceptual*. The interactions between the entities with conceptual semantics, interactions between complex patterns of activity, are not at all simple. Interactions at the level of activity patterns are not directly described by the formal definition of a subsymbolic model; they must be computed by the analyst. Typically, these interactions can be computed only approximately. In other words, there will generally be no precisely valid, complete, computable formal principles at the conceptual level; such principles exist only at the level of individual units – the *subconceptual level*.

(8) c. The subconceptual level hypothesis:
Complete, formal, and precise descriptions of the intu-

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

itive processor are generally tractable not at the conceptual level, but only at the subconceptual level.

In (8c), the qualification "complete, formal, and precise" is important: Conceptual-level descriptions of the intuitive processor's performance can be derived from the subconceptual description, but, unlike the description at the subconceptual level, the conceptual-level descriptions will be either incomplete (describing only certain aspects of the processing) or informal (describing complex behaviors in, say, qualitative terms) or imprecise (describing the performance up to certain approximations or idealizations such as "competence" idealizations away from actual performance). Explicit examples of each of these kinds of conceptual-level descriptions of subsymbolic systems will be considered in Section 9.

Hypotheses (8a–c) can be summarized as:

(8) The subsymbolic hypothesis:
The intuitive processor is a subconceptual connectionist dynamical system that does not admit a complete, formal, and precise conceptual-level description.

This hypothesis is the cornerstone of the subsymbolic paradigm.[4]

2.4. The incompatibility of the symbolic and subsymbolic paradigms. I will now show that the symbolic and subsymbolic paradigms, as formulated above, are incompatible – that hypotheses (4) and (8) about the syntax and semantics of the intuitive processor are not mutually consistent. This issue requires care, because it is well known that one virtual machine can often be implemented in another, that a program written for one machine can be translated into a program for the other. The attempt to distinguish subsymbolic and symbolic computation might well be futile if each can simulate the other. After all, a digital computer is in reality some sort of dynamical system simulating a von Neumann automaton, and in turn, digital computers are usually used to simulate connectionist models. Thus it seems possible that the symbolic and subsymbolic hypotheses (4) and (8) are *both* correct: The intuitive processor can be regarded as a virtual machine for sequentially interpreting rules on one level *and* as a connectionist machine on a lower level.

This possibility fits comfortably within the symbolic paradigm, under a formulation such as:

(10) Valid connectionist models are merely implementations, for a certain kind of parallel hardware, of symbolic programs that provide exact and complete accounts of behavior at the conceptual level. (*To be rejected.*)

However (10) contradicts hypothesis (8c), and is thus incompatible with the subsymbolic paradigm. The symbolic programs that (4) hypothesizes for the intuitive processor could indeed be translated for a connectionist machine; but the translated programs would *not* be the kind of subsymbolic program that (8) hypothesizes. If (10) is correct, (8) is wrong; at the very least, (8c) would have to be removed from the defining hypothesis of the subsymbolic paradigm, weakening it to the point that connectionist modeling does become mere implementation. Such an outcome would constitute a genuine defeat of a research program that I believe many connectionists are pursuing.

What about the reverse relationship, where a symbolic program is used to implement a subsymbolic system? Here it is crucial to realize that the symbols in such programs represent the activation values of units and the strengths of connections. By hypothesis (8b), these do not have conceptual semantics, and thus hypothesis (4b) is violated. The subsymbolic programs that (8) hypothesizes for the intuitive processor can be translated for a von Neumann machine, but the translated programs are *not* the kind of symbolic program that (4) hypothesizes.

These arguments show that unless the hypotheses of the symbolic and subsymbolic paradigms are formulated with some care, the substance of the scientific issue at stake can easily be missed. It is well known that von Neumann machines and connectionist networks can simulate each other. This fact leads some people to adopt the position that the connectionist approach cannot offer anything fundamentally new because we already have Turing machines and, following Church's Thesis, reason to believe that, when it comes to computation, Turing machines are everything. This position, however, mistakes the issue for cognitive science to be the purely syntactic question of whether mental programs are written for Turing/von Neumann machines or connectionist machines. This is a nonissue. If one cavalierly characterizes the two approaches *only syntactically*, using (4a) and (8a) alone, then indeed the issue – connectionist or not connectionist – appears to be "one of AI's wonderful red herrings."[5]

It is a mistake to claim that the connectionist approach has nothing new to offer cognitive science. The issue at stake is a central one: Does the complete formal account of cognition lie at the conceptual level? The position taken by the subsymbolic paradigm is: No – it lies at the subconceptual level.

3. Representation at the subconceptual level

Having hypothesized the existence of a subconceptual level, we must now consider its nature. Hypothesis (8b) leaves open important questions about the semantics of subsymbolic systems. What kind of subconceptual features do the units in the intuitive processor represent? Which activity patterns actually correspond to particular concepts or elements of the problem domain?

There are no systematic or general answers to these questions at the present time; seeking answers is one of the principal tasks for the subsymbolic research paradigm. At present, each individual subsymbolic model adopts particular procedures for relating patterns of activity – activity vectors – to the conceptual-level descriptions of inputs and outputs that define the model's task. The vectors chosen are often values of fine-grained features of the inputs and outputs, based on some preexisting theoretical analysis of the domain. For example, for the task studied by Rumelhart and McClelland (1986), transforming root phonetic forms of English verbs to their past-tense forms, the input and output phonetic strings are represented as vectors of values for context-dependent binary phonetic features. The task description at the conceptual level involves consciously available concepts such as the words "go" and "went," while the subconceptual level used by the model involves a very large number of fine-grained features such as "roundedness preceded

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

by frontalness and followed by backness." The representation of "go" is a large pattern of activity over these features.

Substantive progress in subsymbolic cognitive science requires that systematic commitments be made to vectorial representations for individual cognitive domains. It is important to develop mathematical or empirical methodologies that can adequately constrain these commitments. The vectors chosen to represent inputs and outputs crucially affect a model's predictions, since the generalizations the model makes are largely determined by the similarity structure of the chosen vectors. Unlike symbolic tokens, these vectors lie in a topological space in which some are close together and others far apart.

What kinds of methodologies might be used to constrain the representation at the subconceptual level? The methodology used by Rumelhart and McClelland (1986) in the past-tense model is one that has been fairly widely practiced, particularly in models of language processing: Representational features are borrowed from existing theoretical analyses of the domain and adapted (generally in somewhat ad hoc ways) to meet the needs of connectionist modeling. This methodology clearly renders the subsymbolic approach dependent on other research paradigms in the cognitive sciences and suggests that, certainly in the short term, the subsymbolic paradigm cannot *replace* these other research paradigms. (This is a theme I will return to in the conclusion of the paper.)

A second possible theoretical methodology for studying subconceptual representation relates to the learning procedures that can train hidden units in connectionist networks. Hidden units support internal representations of elements of the problem domain, and networks that train their hidden units are in effect learning effective subconceptual representations of the domain. If we can analyze the representations that such networks develop, we can perhaps obtain principles of subconceptual representation for various problem domains.

A third class of methodology views the task of constraining subconceptual models as the calibration of connectionist models to the human cognitive system. The problem is to determine what vectors should be assigned to represent various aspects of the domain so that the resulting behavior of the connectionist model matches human behavior. Powerful mathematical tools are needed for relating the overall behavior of the network to the choice of representational vectors; ideally, these tools should allow us to *invert* the mapping from representations to behavior so that by starting with a mass of data on human performance we can turn a mathematical crank and have representational vectors pop out. An example of this general type of tool is the technique of *multidimensional scaling* (Shepard 1962), which allows data on human judgments of the similarity between pairs of items in some set to be turned into vectors for representing those items (in a sense). The subsymbolic paradigm needs tools such as a version of multidimensional scaling based on a connectionist model of the process of producing similarity judgments.

Each of these methodologies poses serious research challenges. Most of these challenges are currently being pursued, so far with at best modest success. In the first approach, systematic principles must be developed for adapting to the connectionist context the featural analyses of domains that have emerged from traditional, nonconnectionist paradigms. These principles must reflect fundamental properties of connectionist computation, for otherwise, the hypothesis of connectionist computation is doing no work in the study of mental representation. In the second methodology, principles must be discovered for the representations learned by hidden units, and in the third methodology, principles must be worked out for relating choices of representational vectors to overall system behavior. These are challenging mathematical problems on which the ultimate success of the subsymbolic paradigm rests. Sections 8 and 9 discuss some results related to these mathematical problems, but they are far from strong enough to carry the necessary weight.

The next two sections discuss the relation between the subconceptual level and other levels: The relation to the neural levels is addressed in Section 4, and the relation to the conceptual level is taken up in Section 5.

4. The subconceptual and neural levels

The discussion in the preceding section overlooks an obvious methodology for constraining subconceptual representations – just look at how the brain does it. This brings us back to the parenthetical comment in (7) and the general issue of the relation between the subconceptual and neural levels.[6]

The relation between the subconceptual and neural levels can be addressed in both syntactic and semantic terms. The semantic question is the one just raised: How do representations of cognitive domains as patterns of activity over subconceptual units in the network models of the subsymbolic paradigm relate to representations over neurons in the brain? The syntactic question is: How does the processing architecture adopted by networks in the subsymbolic paradigm relate to the processing architecture of the brain?

There is not really much to say about the semantic question because so little is known about neural representation of higher cognitive domains. When it comes to connectionist modeling of say, language processing, the "just look at how the brain does it" methodology doesn't take one very far towards the goal of constructing a network that does the task at all. Thus it is unavoidable that, for the time being, in subsymbolic models of higher processes, the semantics of network units are much more directly related to conceptual level accounts of these processes than to any neural account. Semantically, the subconceptual level seems at present rather close to the conceptual level, while we have little ground for believing it to be close to the neural level.

This conclusion is at odds with the commonly held view that connectionist models are neural models. That view presumably reflects a bias against semantic considerations in favor of syntactic ones. If one looks only at processing mechanisms, the computation performed by subsymbolic models seems much closer to that of the brain than to that of symbolic models. This suggests that syntactically, the subconceptual level is closer to the neural level than to the conceptual level.

Let us take then the syntactic question: Is the processing architecture adopted by subsymbolic models (8a) well-suited for describing processing at the neural level?

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

Table 1 presents some of the relations between the architectures. The left column lists currently plausible features of some of the most general aspects of the neural architecture, considered at the level of neurons (Crick & Asanuma 1986). The right column lists the corresponding architectural features of the connectionist dynamical systems typically used in subsymbolic models. In the center column, each hit has been indicated by a + and each miss by a −.

In Table 1 the loose correspondence assumed is between neurons and units, between synapses and connections. It is not clear how to make this correspondence precise. Does the activity of a unit correspond to the membrane potential at the cell body? Or the time-averaged firing rate of the neuron? Or the population-averaged firing rate of many neurons? Since the integration of signals between dendritic trees is probably more like the linear integration appearing in quasilinear dynamical systems than is the integration of synaptic signals on a dendrite, would it not be better to view a connection not as an individual synaptic contact but rather as an aggregate contact on an entire dendritic tree?

Table 1. *Relations between the neural and subsymbolic architectures*

Cerebral cortex		**Connectionist dynamical systems**
State defined by continuous numerical variables (potentials, synaptic areas, . . .)	+	State defined by continuous numerical variables (activations, connection strengths)
State variables change continuously in time	+	State variables change continuously in time
Interneuron interaction parameters changeable; seat of knowledge	+	Interunit interaction parameters changeable; seat of knowledge
Huge number of state variables	+	Large number of state variables
High interactional complexity (highly nonhomogeneous interactions)	+	High interactional complexity (highly nonhomogeneous interactions)
Neurons located in 2+1-d space	−	Units have no spatial location
have dense connectivity to nearby neurons;	−	uniformly dense
have geometrically mapped connectivity to distant neurons	−	connections
Synapses located in 3-d space; locations strongly affect signal interactions	−	Connections have no spatial location
Distal projections between areas have intricate topology	−	Distal projections between node pools have simple topology
Distal interactions mediated by discrete signals	−	All interactions nondiscrete
Intricate signal integration at single neuron	−	Signal integration is linear
Numerous signal types	−	Single signal type

Given the difficulty of precisely stating the neural counterpart of components of subsymbolic models, and given the significant number of misses, even in the very general properties considered in Table 1, it seems advisable to keep the question open of the detailed relation between cognitive descriptions at the subconceptual and neural levels. There seems no denying, however, that the subconceptual level is significantly closer to the neural level than is the conceptual level: Symbolic models possess even fewer similarities with the brain than those indicated in Table 1.

The subconceptual level ignores a great number of features of the neural level that are probably extremely important to understanding how the brain computes. Nonetheless, the subconceptual level does incorporate a number of features of neural computation that are almost certainly extremely important to understanding how the brain computes. The general principles of computation at the subconceptual level – computation in high-dimensional, high-complexity dynamical systems – *must* apply to computation in the brain; these principles are likely to be necessary, if not sufficient, to understand neural computation. And while subconceptual principles are not unambiguously and immediately applicable to neural systems, they are certainly more readily applicable than the principles of symbolic computation.

In sum:

(11) The fundamental level of the subsymbolic paradigm, the subconceptual level, lies between the neural and conceptual levels.

As stated earlier, on semantic measures, the subsymbolic level seems closer to the conceptual level, whereas on syntactic measures, it seems closer to the neural level. It remains to be seen whether, as the subsymbolic paradigm develops, this situation will sort itself out. Mathematical techniques like those discussed in the previous section may yield insights into subsymbolic representation that will increase the semantic distance between the subconceptual and conceptual levels. There are already significant indications that as new insights into subsymbolic computation are emerging, and additional information processing power is being added to subsymbolic models, the syntactic distance between the subconceptual and neural levels is increasing. In the drive for more computational power, architectural decisions seem to be driven more and more by mathematical considerations and less and less by neural ones.[7]

Once (11) is accepted, the proper place of subsymbolic models in cognitive science will be clarified. It is common to hear dismissals of a particular subsymbolic model because it is not immediately apparent how to implement it precisely in neural hardware, or because certain neural features are absent from the model. We can now identify two fallacies in such a dismissal. First, following (11): Subsymbolic models should not be viewed as neural models. If the subsymbolic paradigm proves valid, the best subsymbolic models of a cognitive process should one day be shown to be some reasonable higher-level approximation to the neural system supporting that pro-

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

cess. This provides a heuristic that favors subsymbolic models that seem more likely to be reducible to the neural level. But this heuristic is an extremely weak one given how difficult such a judgment must be with the current confusion about the precise neural correlates of units and connections, and the current state of both empirical and theoretical neuroscience.

The second fallacy in dismissing a particular subsymbolic model because of neural unfaithfulness rests on a failure to recognize the role of individual models in the subsymbolic paradigm. A model can make a valuable contribution by providing evidence for general principles that are characteristic of a broad class of subsymbolic systems. The potential value of "ablation" studies of the NETtalk text-to-speech system (Sejnowski & Rosenberg 1986), for example, does not depend entirely on the neural faithfulness of the model, or even on its psychological faithfulness. NETtalk is a subsymbolic system that performs a complex task. What happens to its performance when internal parts are damaged? This provides a significant clue to the general principles of degradation in *all* complex subsymbolic systems: Principles that will apply to future systems that are more faithful as models.

There are, of course, many neural models that do take many of the constraints of neural organization seriously, and for which the analogue of Table 1 would show nearly all hits. But we are concerned here with connectionist models for performing cognitive tasks, and these models typically possess the features displayed in Table 1, with perhaps one or two deviations. The claim is not that neural models don't exist, but rather that they should not be confused with subsymbolic models.

Why is it that neural models of cognitive processes are, generally speaking, currently not feasible? The problem is not an insufficient quantity of data about the brain. The problem, it seems, is that the data are generally of the wrong kind for cognitive modeling. Our information about the nervous system tends to describe its structure, not its dynamic behavior. Subsymbolic systems are dynamical systems with certain kinds of differential equations governing their dynamics. If we knew which dynamical variables in the neural system for some cognitive task were the critical ones for performing that task, and what the "equations of motion" were for those variables, we could use that information to build neurally faithful cognitive models. But generally what we know instead are endless static properties of how the hardware is arranged. Without knowing which (if any) of these structures support relevant dynamical processes, and what equations govern those processes, we are in a position comparable to someone attempting to model the solar system, armed with voluminous data on the colored bands of the planets but with no knowledge of Newton's Laws.

To summarize:

(12) a. Unlike the symbolic architecture, the subsymbolic architecture possesses a number of the most general features of the neural architecture.
b. However, the subsymbolic architecture lacks a number of the more detailed but still quite general features of the neural architecture; the subconceptual level of analysis is higher than the neural level.
c. For most cognitive functions, neuroscience cannot provide the relevant information to specify a cognitive model at the neural level.
d. The general cognitive principles of the subconceptual level will probably be important contributors to future discoveries of those specifications of neural computations that we now lack.

5. Reduction of cognition to the subconceptual level

The previous section considered the relationship between the fundamental level of the subsymbolic paradigm – the subconceptual level – and the neural level. The remainder of this article will focus on relations between the subconceptual and conceptual levels; these have so far only been touched upon briefly (in (8c)). Before proceeding, however, it is worth summarizing the relationships between the levels, including those that will be discussed in the remainder of the article.

Imagine three physical systems: a brain that is executing some cognitive process, a massively parallel connectionist computer running a subsymbolic model of that process, and a von Neumann computer running a symbolic model of the same process. The cognitive process may involve conscious rule application, intuition, or a combination of the two. According to the subsymbolic paradigm, here are the relationships:

(13) a. Describing the brain at the neural level gives a neural model.
b. Describing the brain approximately, at a higher level – the subconceptual level – yields, to a good approximation, the model running on the connectionist computer, when it too is described at the subconceptual level. (At this point, this is a goal for future research. It could turn out that the degree of approximation here is only rough; this would still be consistent with the subsymbolic paradigm.)
c. We can try to describe the connectionist computer at a higher level – the conceptual level – by using the patterns of activity that have conceptual semantics. If the cognitive process being executed is conscious rule application, we will be able to carry out this conceptual-level analysis with reasonable precision, and will end up with a description that closely matches the symbolic computer program running on the von Neumann machine.
d. If the process being executed is an intuitive process, we will be unable to carry out the conceptual-level description of the connectionist machine precisely. Nonetheless, we will be able to produce various approximate conceptual-level descriptions that correspond to the symbolic computer program running on the von Neumann machine in various ways.

For a cognitive process involving both intuition and conscious rule application, (13c) and (13d) will each apply to certain aspects of the process.

The relationships (13a) and (13b) were discussed in the previous section. The relationship (13c) between a subsymbolic implementation of the conscious rule interpreter and a symbolic implementation is discussed in Section 6. The relations (13d) between subsymbolic and symbolic accounts of intuitive processing are considered in Section 9. These relations hinge on certain subsymbolic computa-

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

Table 2. *Three cognitive systems and three levels of description*

Level	(process)	Cognitive system: Brain		Subsymbolic		Symbolic
Conceptual	(intuition)	?		rough approximation	~	exact
	(conscious rule application)	?		good approximation	≃	exact
Subconceptual		good approximation	≃	exact		
Neural		exact				

tional principles operative at the subconceptual level (13b); these are briefly discussed in Section 8. These principles are of a new kind for cognitive science, giving rise to the foundational considerations taken up in Section 7.

The relationships in (13) can be more clearly understood by reintroducing the concept of "virtual machine." If we take one of the three physical systems and describe its processing at a certain level of analysis, we get a virtual machine that I will denote "$\text{system}_{\text{level}}$". Then (13) can be written:

(14) a. $\text{brain}_{\text{neural}}$ = neural model
b. $\text{brain}_{\text{subconceptual}} \simeq \text{connectionist}_{\text{subconceptual}}$
c. $\text{connectionist}_{\text{conceptual}} \simeq \text{von Neumann}_{\text{conceptual}}$ (conscious rule application)
d. $\text{connectionist}_{\text{conceptual}} \sim \text{von Neumann}_{\text{conceptual}}$ (intuition)

Here, the symbol ≃ means "equals to a good approximation" and ~ means "equals to a crude approximation." The two nearly equal virtual machines in (14c) both describe what I have been calling the "conscious rule interpreter." The two roughly similar virtual machines in (14d) provide the two paradigms' descriptions of the intuitive processor at the conceptual level.

Table 2 indicates these relationships and also the degree of exactness to which each system can be described at each level – the degree of precision to which each virtual machine is defined. The levels included in Table 2 are those relevant to predicting high-level behavior. Of course each system can also be described at lower levels, all the way down to elementary particles. However, levels below an exactly describable level can be ignored from the point of view of predicting high-level behavior, since it is possible (in principle) to do the prediction at the highest level that can be exactly described (it is presumably much harder to do the same at lower levels). This is why in the symbolic paradigm any descriptions below the conceptual level are not viewed as significant. For modeling high-level behavior, how the symbol manipulation happens to be implemented can be ignored – it is not a relevant part of the cognitive model. In a subsymbolic model, exact behavioral prediction must be performed at the subconceptual level, but how the units happen to be implemented is not relevant.

The relation between the conceptual level and lower levels is fundamentally different in the subsymbolic and symbolic paradigms. This leads to important differences in the kind of explanations the paradigms offer of conceptual-level behavior, and the kind of reduction used in these explanations. A symbolic model is a *system* of interacting processes, all with the same conceptual-level semantics as the task behavior being explained. Adopting the terminology of Haugeland (1978), this *systematic explanation* relies on a *systematic reduction* of the behavior that involves no shift of semantic domain or *dimension*. Thus a game-playing program is composed of subprograms that generate possible moves, evaluate them, and so on. In the symbolic paradigm, these systematic reductions play the major role in explanation. The lowest-level processes in the systematic reduction, still with the original semantics of the task domain, are then themselves reduced by *intentional instantiation*: they are implemented exactly by other processes with different semantics but the same form. Thus a move-generation subprogram with game semantics is instantiated in a system of programs with list-manipulating semantics. This intentional instantiation typically plays a minor role in the overall explanation, if indeed it is regarded as a cognitively relevant part of the model at all.

Thus cognitive explanations in the symbolic paradigm rely primarily on reductions involving no dimensional shift. This feature is not shared by the subsymbolic paradigm, where accurate explanations of intuitive behavior require descending to the subconceptual level. The elements in this explanation, the units, do *not* have the semantics of the original behavior: that is the content of the subconceptual unit hypothesis, (8b). In other words:

(15) Unlike symbolic explanations, subsymbolic explanations rely crucially on a semantic ("dimensional") shift that accompanies the shift from the conceptual to the subconceptual levels.

The overall dispositions of cognitive systems are explained in the subsymbolic paradigm as approximate higher-level regularities that emerge from quantitative laws operating at a more fundamental level with different semantics. This is the kind of reduction familiar in natural science, exemplified by the explanation of the laws of thermodynamics through a reduction to mechanics that involves shifting the dimension from thermal semantics to molecular semantics. (Section 9 discusses some explicit subsymbolic reductions of symbolic explanatory constructs.)

Indeed the subsymbolic paradigm repeals the other features that Haugeland identified as newly introduced into scientific explanation by the symbolic paradigm. The inputs and outputs of the system are not quasilinguistic representations but good old-fashioned numerical vectors. These inputs and outputs have semantic interpretations, but these are not constructed recursively from interpretations of embedded constitu-

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

ents. The fundamental laws are good old-fashioned numerical equations.

Haugeland went to considerable effort to legitimize the form of explanation and reduction used in the symbolic paradigm. The explanations and reductions of the subsymbolic paradigm, by contrast, are of a type well-established in natural science.

In summary, let me emphasize that in the subsymbolic paradigm, the conceptual and subconceptual levels are not related as the levels of a von Neumann computer (high-level-language program, compiled low-level program, etc.). The relationship between subsymbolic and symbolic models is more like that between quantum and classical mechanics. Subsymbolic models accurately describe the microstructure of cognition, whereas symbolic models provide an approximate description of the macrostructure. An important job of subsymbolic theory is to delineate the situations and the respects in which the symbolic approximation is valid, and to explain why.

6. Conscious rule application in the subsymbolic paradigm

In the symbolic paradigm, both conscious rule application and intuition are described at the conceptual level; that is, conscious and unconscious rule interpretation, respectively. In the subsymbolic paradigm, conscious rule application can be formalized in the conceptual level but intuition must be formalized at the subconceptual level. This suggests that a subsymbolic model of a cognitive process that involves both intuition and conscious rule interpretation would consist of two components using quite different formalisms. While this hybrid formalism might have considerable practical value, there are some theoretical problems with it. How would the two formalisms communicate? How would the hybrid system evolve with experience, reflecting the development of intuition and the subsequent remission of conscious rule application? How would the hybrid system elucidate the fallibility of actual human rule application (e.g., logic)? How would the hybrid system get us closer to understanding how conscious rule application is achieved neurally?

All these problems can be addressed by adopting a unified subconceptual-level analysis of both intuition and conscious rule interpretation. The virtual machine that is the conscious rule interpreter is to be implemented in a lower-level virtual machine: the same connectionist dynamical system that models the intuitive processor. How this can, in principle, be achieved is the subject of this section. The relative advantages and disadvantages of implementing the rule interpreter in a connectionist dynamical system, rather than a von Neumann machine, will also be considered.

Section 2.1 described the power of natural language for the propagation of cultural knowledge and the instruction of novices. Someone who has mastered a natural language has a powerful trick available for performing in domains where experience has been insufficient for the development of intuition: Verbally expressed rules, whether resident in memory or on paper, can be used to direct a step-by-step course to an answer. Once subsymbolic models have achieved a sufficient subset of the power to process natural language, they will be able to exploit the same trick. A subsymbolic system with natural language competence will be able to encode linguistic expressions as patterns of activity; like all other patterns of activity, these can be stored in connectionist memories using standard procedures. If the linguistic expressions stored in memory happen to be rules, the subsymbolic system can use them to solve problems sequentially in the following way. Suppose, for concreteness, that the rules stored in memory are production rules of the form "if *condition* holds, then do *action*." If the system finds itself in a particular situation where *condition* holds, then the stored production can be retrieved from the connectionist memory via the characteristic *content-addressability* of these memories: of the activity pattern representing the entire production, the subpart that pertains to *condition* is present, and this then leads to the reinstantiation in the memory of the entire pattern representing the production. The competence of the subsymbolic system to process natural language must include the ability to take the portion of the reinstantiated pattern that encodes the verbal description of *action*, and actually execute the action it describes; that is, the subsymbolic system must be able to *interpret*, in the computational sense of the term, the memorized description of *action*. The result is a subsymbolic implementation of a production system, built purely out of subsymbolic natural language processing mechanisms. A connectionist account of natural language processes must eventually be developed as part of the subsymbolic paradigm, because natural language processes of fluent speakers are intuitive and thus, according to the subsymbolic hypothesis (8), must be modeled at the subconceptual level using subsymbolic computation.

In summary:

(16) The competence to represent and process linguistic structures in a native language is a competence of the human intuitive processor; the subsymbolic paradigm assumes that this competence can be modeled in a subconceptual connectionist dynamical system. By combining such linguistic competence with the memory capabilities of connectionist systems, sequential rule interpretation can be implemented.

Now note that our subsymbolic system can use its stored rules to perform the task. The standard learning procedures of connectionist models now turn this experience of performing the task into a set of weights for going from inputs to outputs. Eventually, after enough experience, the task can be performed directly by these weights. The input activity generates the output activity so quickly that before the relatively slow rule-interpretation process has a chance to reinstantiate the first rule in memory and interpret it, the task is done. With intermediate amounts of experience, some of the weights are well enough in place to prevent some of the rules from having the chance to instantiate, while others are not, enabling other rules to be retrieved and interpreted.

6.1. Rule interpretation, consciousness, and seriality. What about the conscious aspect of rule interpretation? Since consciousness seems to be a quite high-level description of mental activity, it is reasonable to suspect that it

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

reflects the very coarse structure of the cognitive dynamical system. This suggests the following hypothesis:

(17) The contents of consciousness reflect only the large-scale structure of activity patterns: subpatterns of activity that are extended over spatially large regions of the network and that are stable for relatively long periods of time.

(See Rumelhart, Smolensky, McClelland & Hinton 1986. Note that (17) hypothesizes a *necessary* – not a *sufficient* – condition for an aspect of the subsymbolic state to be relevant to the conscious state.) The spatial aspect of this hypothesis has already played a major role in this article – it is in fact a restatement of the subconceptual unit hypothesis, (8b): Concepts that are consciously accessible correspond to patterns over large numbers of units. It is the temporal aspect of hypothesis (17) that is relevant here. The rule interpretation process requires that the retrieved linguistically coded rule be maintained in memory while it is being interpreted. Thus the pattern of activity representing the rule must be stable for a relatively long time. In contrast, after connections have been developed to perform the task directly, there is no correspondingly stable pattern formed during the performance of the task. Thus the loss of conscious phenomenology with expertise can be understood naturally.

On this account, the sequentiality of the rule interpretation process is not built into the architecture; rather, it is linked to our ability to follow only one verbal instruction at a time. Connectionist memories have the ability to retrieve a single stored item, and here this ability is called upon so that the linguistic interpreter is not required to interpret multiple instructions simultaneously.

It is interesting to note that the preceding analysis also applies to nonlinguistic rules: Any notational system that can be appropriately interpreted will do. For example, another type of rule might be a short series of musical pitches; a memorized collection of such rules would allow a musician to play a tune by conscious rule interpretation. With practice, the need for conscious control goes away. Since pianists learn to interpret several notes simultaneously, the present account suggests that a pianist might be able to apply more than one musical rule at a time; if the pianist's memory for these rules can simultaneously recall more than one, it would be possible to generate multiple musical lines simultaneously using conscious rule interpretation. A symbolic account of such a process would involve something like a production system capable of firing multiple productions simultaneously.

Finally, it should be noted that even if the memorized rules are assumed to be linguistically coded, the preceding analysis is uncommitted about the form the encoded rules take in memory: phonological, orthographic, semantic, or whatever.

6.2. Symbolic versus subsymbolic implementation of rule interpretation. The (approximate) implementation of the conscious rule interpreter in a subsymbolic system has both advantages and disadvantages relative to an (exact) implementation in a von Neumann machine.

The main disadvantage is that subconceptual representation and interpretation of linguistic instructions is very difficult and we are not actually able to do it now. Most existing subsymbolic systems simply don't use rule interpretation.[8] Thus they miss out on all the advantages listed in (2). They can't take advantage of rules to check the results produced by the intuitive processor. They can't bootstrap their way into a new domain using rules to generate their own experience: they must have a teacher generate it for them.[9]

There are several advantages of a subconceptually implemented rule interpreter. The intuitive processor and rule interpreter are highly integrated, with broadband communication between them. Understanding how this communication works should allow the design of efficient hybrid symbolic/subsymbolic systems with effective communication between the processors. A principled basis is provided for studying how rule-based knowledge leads to intuitive knowledge. Perhaps most interesting, in a subsymbolic rule interpreter, the process of rule selection is intuitive! Which rule is reinstantiated in memory at a given time is the result of the associative retrieval process, which has many nice properties. The best match to the productions' conditions is quickly computed, and even if no match is very good, a rule can be retrieved. The selection process can be quite context-sensitive.

An integrated subsymbolic rule interpreter/intuitive processor in principle offers the advantages of both kinds of processing. Imagine such a system creating a mathematical proof. The intuitive processor would generate goals and steps, and the rule interpreter would verify their validity. The serial search through the space of possible steps, which is necessary in a purely symbolic approach, is replaced by the intuitive generation of possibilities. Yet the precise adherence to strict inference rules that is demanded by the task can be enforced by the rule interpreter; the creativity of intuition can be exploited while its unreliability can be controlled.

6.3. Two kinds of knowledge – one knowledge medium. Most existing subsymbolic systems perform tasks without serial rule interpretation: Patterns of activity representing inputs are directly transformed (possibly through multiple layers of units) into patterns of activity representing outputs. The connections that mediate this transformation represent a form of task knowledge that can be applied with massive parallelism: I will call it *P-knowledge*. For example, the P-knowledge in a native speaker presumably encodes lexical, morphological, syntactic, semantic, and pragmatic constraints in such a form that all these constraints can be satisfied in parallel during comprehension and generation.

The connectionist implementation of sequential rule interpretation described above displays a second form that knowledge can take in a subsymbolic system. The stored activity patterns that represent rules also constitute task knowledge: Call it *S-knowledge*. Like P-knowledge, S-knowledge is embedded in connections: the connections that enable part of a rule to reinstantiate the entire rule. Unlike P-knowledge, S-knowledge cannot be used with massive parallelism. For example, a novice speaker of some language cannot satisfy the constraints contained in two memorized rules simultaneously; they must be serially reinstantiated as patterns of

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

activity and separately interpreted. Of course, the connections responsible for reinstantiating these memories operate in parallel, and indeed these connections contain within them the potential to reinstantiate either of the two memorized rules. But these connections are so arranged that only one rule at a time can be reinstantiated. The retrieval of each rule is a parallel process, but the satisfaction of the constraints contained within the two rules is a serial process. After considerable experience, P-knowledge is created: connections that can simultaneously satisfy the constraints represented by the two rules.

P-knowledge is considerably more difficult to create than S-knowledge. To encode a constraint in connections so that it can be satisfied in parallel with thousands of others is not an easy task. Such an encoding can only be learned through considerable experience in which that constraint has appeared in many different contexts, so that the connections enforcing the constraint can be tuned to operate in parallel with those enforcing a wide variety of other constraints. S-knowledge can be acquired (once the linguistic skills on which it depends have been encoded into P-knowledge, of course) much more rapidly. For example, simply reciting a verbal rule over and over will usually suffice to store it in memory, at least temporarily.

That P-knowledge is so highly context-dependent while the rules of S-knowledge are essentially context-independent is an important computational fact underlying many of the psychological explanations offered by subsymbolic models. Consider, for example, Rumelhart and McClelland's (1986) model of the U-shaped curve for past-tense production in children. The phenomenon is striking: A child is observed using *goed* and *wented* when at a much younger age *went* was reliably used. This is surprising because we are prone to think that such linguistic abilities rest on knowledge that is encoded in some context-independent form such as "the past tense of *go* is *went*." Why should a child *lose* such a rule once acquired? A traditional answer invokes the acquisition of a different context-independent rule, such as "the past tense of *x* is *x* + *ed*" which, for one reason or another, takes precedence. The point here, however, is that there is nothing at all surprising about the phenomenon when the underlying knowledge is assumed to be context-dependent and not context-independent. The young child has a small vocabulary of largely irregular verbs. The connections that implement this P-knowledge are reliable in producing the large pattern of activity representing *went*, as well as those representing a small number of other past-tense forms. Informally we can say that the connections producing *went* do so in the context of the other vocabulary items that are also stored in the same connections. There is no guarantee that these connections will produce *went* in the context of a different vocabulary. As the child acquires additional vocabulary items, most of which are regular, the context radically changes. Connections that were, so to speak, perfectly adequate for creating *went* in the old context now have to work in a context where very strong connections are trying to create forms ending in *-ed*; the old connections are not up to the new task. Only through extensive experience trying to produce *went* in the new context of many regular verbs can the old connections be modified to work in the new context. In particular, strong new connections must be added that, when the input pattern encodes *go*, cancel the *-ed* in the output; these were not needed before.

These observations about context-dependence can also be framed in terms of inference. If we choose to regard the child as using knowledge to infer the correct answer *went*, then we can say that after the child has added more knowledge (about new verbs), the ability to make the (correct) inference is lost. In this sense the child's inference process is nonmonotonic – perhaps this is why we find the phenomenon surprising. As will be discussed in Section 8, nonmonotonicity is a fundamental property of subsymbolic inference.

To summarize:

(18) a. Knowledge in subsymbolic systems can take two forms, both resident in the connections.
b. The knowledge used by the conscious rule interpreter lies in connections that reinstantiate patterns encoding rules; task constraints are coded in context-independent rules and satisfied serially.
c. The knowledge used in intuitive processing lies in connections that constitute highly context-dependent encodings of task constraints that can be satisfied with massive parallelism.
d. Learning such encodings requires much experience.

7. Subsymbolic definition of cognitive systems and some foundational issues

In order for the subconceptual level to be rightly viewed as a level for practicing cognitive science, it is necessary that the principles formulated at this level truly be principles of cognition. Since subsymbolic principles are neither conceptual-level nor neural-level principles, it is not immediately apparent what kind of cognitive principles they might be. The structure of subsymbolic models is that of a dynamical system; in what sense do these models embody principles of cognition rather than principles of physics?

What distinguishes those dynamical systems that are cognitive from those that are not? At this point the types of dynamical systems being studied in connectionist cognitive science lack anything that could justly be called an intentional psychology. In this section I wish to show that it is nonetheless possible to distinguish the sort of dynamical systems that have so far been the object of study in connectionist cognitive science from the dynamical systems that have traditionally been the subject matter of physics, and that the questions being studied are indeed questions of cognition.

A crucial property of cognitive systems broadly construed is that over a wide variety of environments they can maintain, at an adequately constant level, the degree to which a significant number of *goal conditions* are met. Here I intend the teleological, rather than the intentional, sense of "goal." A river, for example, is a complex dynamical system that responds sensitively to its environment – but about the only condition that it can satisfy over a large range of environments is going downhill. A cockroach manages, over an annoyingly extensive range of environments, to maintain its nutritive intake, its reproductive demands, its oxygen intake, even its probability of getting smashed, all within a relatively narrow

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

band. The repertoire of conditions that people can keep satisfied, and the range of environments under which this relative constancy can be maintained, provides a measure worthy of the human cognitive capacity.

(19) Cognitive system:
A necessary condition for a dynamical system to be *cognitive* is that, under a wide variety of environmental conditions, it maintains a large number of goal conditions. The greater the repertoire of goals and variety of tolerable environmental conditions, the greater the cognitive capacity of the system.

The issue of complexity is crucial here. A river (or a thermostat) only fails to be a cognitive dynamical system because it cannot satisfy a *large* range of goals under a *wide* range of conditions.[10] Complexity is largely what distinguishes the dynamical systems studied in the subsymbolic paradigm from those traditionally studied in physics. Connectionist dynamical systems have great complexity: The information content in their weights is very high. Studying the extent to which a connectionist dynamical system can achieve complex goals in complex environments requires grappling with complexity in dynamical systems in a way that is traditionally avoided in physics. In cognitive modeling, many of the basic questions concern the detailed dynamics of a distinct pattern of activation in a system with a particular initial state and a particular set of interaction strengths that are highly nonhomogeneous. This is like asking a physicist: "Suppose we have a gas with 10,000 particles with the following 10,000 different masses and the following 500,000 different forces between them. Suppose we start them at rest in the following 10,000 positions. What are the trajectories of the following 20 particles?" This is indeed a question about a dynamical system, and is, in a sense, a question of physics. It is this kind of question, however, that is avoided at all costs in physics. The physicist we consulted is likely to compute the mean collision times for the particles assuming equal masses, random starting positions, and uniformly random interactions, and say "if that isn't good enough, then take your question to a computer."[11]

Nonetheless, physics has valuable concepts and techniques to contribute to the study of connectionist dynamical systems. Insights from physics have already proved important in various ways in the subsymbolic paradigm (Hinton & Sejnowski 1983a; Sejnowski 1976; Smolensky 1983).

Various subsymbolic models have addressed various goals and environments. A very general goal that is of particular importance is:

(20) *The prediction goal:* Given some partial information about the environmental state, correctly infer missing information.

What is maintained here is the degree of match between predicted values and the actual values for the unknowns. Maintenance of this match over the wide range of conditions found in a complex environment is a difficult task. Special cases of this task include predicting the depth of an object from retinal images, the future location of a moving object, the change in certain aspects of an electric circuit given the changes in other aspects, or the propositions implied by a text. The prediction goal is obviously an important one, because it can serve so many other goals: Accurate prediction of the effects of actions allows the selection of those leading to desired effects.

A closely related goal is:

(21) *The prediction-from-examples goal:* Given more and more examples of states from an environment, achieve the prediction goal with increasing accuracy in that environment.

For the prediction goal we ask: What inference procedures and knowledge about an environment must a dynamical system possess to be able to predict that environment? For the prediction-from-examples goal we go further and ask: What learning procedures must a dynamical system possess to be able to acquire the necessary knowledge about an environment from examples?

The goals of prediction and prediction-from-examples are the subject of many principles of the subsymbolic paradigm. These are indeed cognitive principles. They will be taken up in the next section; first, however, I would like to consider some implications of this characterization of a cognitive system for certain foundational issues: semantics, rationality, and the constituent structure of mental states. It would be absurd to suggest that the following few paragraphs constitute definitive treatments of these issues; the intent is rather to indicate specific points where subsymbolic research touches on these issues and to sow seeds for further analysis.

7.1. Semantics and rationality in the subsymbolic paradigm. The subsymbolic characterization of a cognitive system (19) intrinsically binds cognitive systems both to states of the environment and to goal conditions. It therefore has implications for the question: How do states of a subsymbolic system get their meanings and truth conditions? A starting point for an answer is suggested in the following hypothesis:

(22) Subsymbolic semantics:
A cognitive system adopts various internal states in various environmental conditions. To the extent that the cognitive system meets its goal conditions in various environmental conditions, its internal states are *veridical representations* of the corresponding environmental states, with respect to the given goal conditions.

For the prediction goal, for example, a state of the subsymbolic system is a veridical representation of the current environmental state to the extent that it leads to correct predictions.

According to hypothesis (22), it is not possible to localize a failure of veridical representation. Any particular state is part of a large causal system of states, and failures of the system to meet goal conditions cannot in general be localized to any particular state or state component.[12] In subsymbolic systems, this *assignment of blame problem* (Minsky 1963) is a difficult one, and it makes programming subsymbolic models by hand very tricky. Solving the assignment of blame problem is one of the central accomplishments of the automatic network programming procedures: the learning procedures of the subsymbolic paradigm.

The characterization (19) of cognitive systems relates to rationality as well. How can one build a rational machine?

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

How can internal processes (e.g., inference) be guaranteed to preserve veridical semantic relationships (e.g., be truth preserving)? These questions now become: How can the connection strengths be set so that the subsymbolic system will meet its goal conditions? Again, this is a question answered by the scientific discoveries of the subsymbolic paradigm: particular procedures for programming machines to meet certain goals – especially learning procedures to meet adaptation goals such as prediction-from-examples.

Let me compare this subsymbolic approach to veridicality with a symbolic approach to truth preservation offered by Fodor (1975; 1987). In the context of model-theoretic semantics for a set of symbolic formulae, proof theory provides a set of symbol manipulations (rules of inference) guaranteed to preserve truth conditions. Thus if an agent possesses knowledge in the symbolic form $p \rightarrow q$ and additional knowledge p, then by syntactic operations the agent can produce q; proof theory guarantees that the truth conditions of the agent's knowledge (or beliefs) has not changed.

There are fairly direct subsymbolic counterparts to this proof theoretic account. The role of logical inference is played by statistical inference. By explicitly formalizing tasks like prediction as statistical inference tasks, it is possible to prove for appropriate systems that subsymbolic computation is valid in a sense directly comparable to symbolic proof. Further discussion of this point, which will appear in Section 9.1, must await further examination of the computational framework of the subsymbolic paradigm, which is the subject of Section 8.

Note that the proof theoretic account explains the tautological inference of q from p and $p \rightarrow q$, but it leaves to an independent module an account of how the agent acquired the knowledge $p \rightarrow q$ that licenses the inference from p to q. In the subsymbolic account, the veridicality problem is tied inextricably to the environment in which the agent is trying to satisfy the goal conditions – subsymbolic semantics is intrinsically situated. The subsymbolic analysis of veridicality involves the following basic questions: How can a cognitive system be put in a novel environment and learn to create veridical internal representations that allow valid inferences about that environment so that goal conditions can be satisfied? How can it pick up information from its environment? These are exactly the questions addressed by subsymbolic learning procedures.

Note that in the subsymbolic case, the internal processing mechanisms (which can appropriately be called inference procedures) do not, of course, directly depend causally on the environmental state that may be internally represented or on the veridicality of that representation. In that sense, they are just as formal as syntactic symbol manipulations. The fact that a subsymbolic system can generate veridical representations of the environment (e.g., make valid predictions) is a result of extracting information from the environment and internally coding it in its weights through a learning procedure.

7.2. Constituent structure of mental states. Fodor and Pylyshyn have argued (e.g., Fodor 1975; Pylyshyn 1984) that mental states must have constituent structure, and they have used this argument against the connectionist approach (Fodor & Pylyshyn 1988). Their argument applies, however, only to ultra-local connectionist models (Ballard & Hayes 1984); it is quite inapplicable to the distributed connectionist systems considered here. A mental state in a subsymbolic system is a pattern of activity with a constituent structure that can be analyzed at both the conceptual and the subconceptual levels. In this section I offer a few general observations on this issue; the connectionist representation of complex structures is an active area of research (Smolensky 1987; Touretzky 1986), and many difficult problems remain to be solved (for futher discussion see Smolensky 1988).

At the conceptual level, a connectionist mental state contains constituent subpatterns that have conceptual interpretations. Pylyshyn, in a debate over the connectionist approach at the 1984 meeting of the Cognitive Science Society, suggested how to extract these conceptual constituents with the following example: The connectionist representation of *coffee* is the representation of *cup with coffee* minus the representation of *cup without coffee*. To carry out this suggestion, imagine a crude but adequate kind of distributed semantic representation, in which the interpretation of *cup with coffee* involves the activity of network units representing features like brown liquid with flat top surface, brown liquid with curved sides and bottom surface, brown liquid contacting porcelain, hot liquid, upright container with a handle, burnt odor, and so forth. We should really use subconceptual features, but even these features are sufficiently low-level to make the point. Following Pylyshyn, we take this representation of the interpretation of *cup with coffee* and subtract from it the representation of the interpretation of *cup without coffee*, leaving the representation of *coffee*. What remains, in fact, is a pattern of activity with active features such as brown liquid with flat top surface, brown liquid with curved sides and bottom surface, brown liquid contacting porcelain, hot liquid, and burnt odor. This represents *coffee*, in some sense – but *coffee in the context of cup*.

In using Pylyshyn's procedure for determining the connectionist representation of *coffee*, there is nothing sacred about starting with *cup with coffee*: why not start with *can with coffee*, *tree with coffee*, or *man with coffee*, and subtract the corresponding representation of *X without coffee*? Thinking back to the distributed featural representation, it is clear that each of these procedures produces quite a different result for "the" connectionist representation of *coffee*. The pattern representing *coffee* in the context of *cup* is quite different from the pattern representing *coffee* in the context of *can*, *tree*, or *man*.

The pattern representing *cup with coffee* can be decomposed into conceptual-level constituents, one for *coffee* and another for *cup*. This decomposition differs in two significant ways from the decomposition of the symbolic expression *cup with coffee*, into the three constituents, *coffee*, *cup*, and *with*. First, the decomposition is quite approximate. The pattern of features representing *cup with coffee* may well, as in the imagined case above, possess a subpattern that can be identified with *coffee*, as well as a subpattern that can be identified with *cup*; but these subpatterns will in general not be defined precisely and there will typically remain features that can be identified only with the interaction of the two (as in brown liquid contacting porcelain). Second, whatever the subpattern identified with *coffee*, unlike the symbol *coffee*, it

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

is a context-dependent constituent, one whose internal structure is heavily influenced by the structure of which it is a part.

These constituent subpatterns representing *coffee* in varying contexts are activity vectors that are not identical, but possess a rich structure of commonalities and differences (a family resemblance, one might say). The commonalities are directly responsible for the common processing implications of the interpretations of these various phrases, so 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.

The conceptual-level constituents of mental states are activity vectors, which themselves have constituent structure at the subconceptual level: the individual units' activities. To summarize the relationship between these notions of constituent structure in the symbolic and subsymbolic paradigms, let's call each *coffee* vector the (connectionist) symbol for coffee in the given context. Then we can say that the context alters the internal structure of the symbol; the activities of the subconceptual units that comprise the symbol – its subsymbols – change across contexts. In the symbolic paradigm, a symbol is effectively contextualized by surrounding it with other symbols in some larger structure. In other words:

(23) Symbols and context dependence:
In the symbolic paradigm, the context of a symbol is manifest around it and consists of other symbols; in the subsymbolic paradigm, the context of a symbol is manifest inside it and consists of subsymbols.

(Compare Hofstadter 1979; 1985.)

8. Computation at the subconceptual level

Hypothesis (8a) offers a brief characterization of the connectionist architecture assumed at the subconceptual level by the subsymbolic paradigm. It is time to bring out the computational principles implicit in that hypothesis.

8.1. Continuity. According to (8a), a connectionist dynamical system has a continuous space of states and changes state continuously in time. I take time in this section to motivate at some length this assumption of continuity, because it plays a central role in the characterization of subsymbolic computation and because readers familiar with the literature on connectionist models will no doubt require that I reconcile the continuity assumption with some salient candidate counterexamples.

Within the symbolic paradigm, the simplest, most straightforward formalizations of a number of cognitive processes have quite discrete characters:

(24) a. Discrete memory locations, in which items are stored without mutual interaction.
b. Discrete memory storage and retrieval operations, in which an entire item is stored or retrieved in a single, atomic (primitive) operation.
c. Discrete learning operations, in which new rules become available for use in an all-or-none fashion.
d. Discrete inference operations, in which conclusions become available for use in an all-or-none fashion.
e. Discrete categories, to which items either belong or do not belong.
f. Discrete production rules, with conditions that are either satisfied or not satisfied, and actions that either execute or do not execute.

These discrete features come "for free" in the symbolic paradigm: Of course, any one of them can be softened but only by explicitly building in machinery to do so.

Obviously (24) is a pretty crude characterization of cognitive behavior. Cognition seems to be a richly interwoven fabric of graded, continuous processes and discrete, all-or-none processes. One way to model this interplay is to posit separate discrete and continuous processors in interaction. Some theoretical problems with this move were mentioned in Section 6, where a unified formalism was advocated. It is difficult to introduce a hard separation between the soft and the hard components of processing. An alternative is to adopt a fundamentally symbolic approach, but to soften various forms of discreteness by hand. For example, the degree of match to conditions of production rules can be given numerical values, productions can be given strengths, interactions between separately stored memory items can be put in by hand, and so on (Anderson 1983).

The subsymbolic paradigm offers another alternative. All the discrete features of (24) are neatly swept aside in one stroke by adopting a continuous framework that applies at the subconceptual level. Then, when the continuous system is analyzed at the higher, conceptual level, various aspects of discreteness emerge naturally and inevitably, without explicit machinery having been devised to create this discreteness. These aspects of "hardness" are intrinsically embedded in a fundamentally "soft" system. The dilemma of accounting for both the hard and soft aspects of cognition is solved by using the passage from a lower level of analysis to a higher level to introduce natural changes in the character of the system: The emergent properties can have a different nature from the fundamental properties. This is the story to be fleshed out in the remainder of the paper. It rests on the fundamental continuity of subsymbolic computation, which is further motivated in the remainder of this section (for further discussion see Smolensky 1988).

It may appear that the continuous nature of subsymbolic systems is contradicted by the fact that it is easy to find in the connectionist literature models that are quite within the spirit of the subsymbolic paradigm, but which have neither continuous state spaces nor continuous dynamics. For example, models having units with binary values that jump discretely on the ticks of a discrete clock (the Boltzmann machine, Ackley et al. 1985; Hinton & Sejnowski 1983a; harmony theory, Smolensky 1983; 1986a). I will now argue that these models should be viewed as discrete simulations of an underlying continuous model, considering first discretization of time and then discretization of the units' values.

Dynamical systems evolving in continuous time are almost always simulated on digital computers by discretizing time. Since subsymbolic models have almost always been simulated on digital computers, it is no surprise that they too have been simulated by discretizing time. The equations defining the dynamics of the models can be understood more easily by most cognitive scien-

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

tists if the differential equations of the underlying continuous dynamical system are avoided in favor of the discrete-time approximations that actually get simulated.

When subsymbolic models use binary-valued units, these values are best viewed not as symbols like *T* and *NIL* that are used for conditional branching tests, but as numbers (not numerals!) like 1 and 0 that are used for numerical operations (e.g., multiplication by weights, summation, exponentiation). These models are formulated in such a way that they are perfectly well-defined for continuous values of the units. Discrete numerical unit values are no more than a simplification that is sometimes convenient.[13]

As historical evidence that underlying subsymbolic models are continuous systems, it is interesting to note that when the theoretical conditions that license the discrete approximation have changed, the models have reverted to continuous values. In the harmony/energy optima model, when the jumpy stochastic search was replaced by a smooth deterministic one (Rumelhart, Smolensky, McClelland & Hinton 1986), the units were changed to continuous ones.[14]

A second, quite dramatic, piece of historical evidence is a case where switching from discrete to continuous units made possible a revolution in subsymbolic learning theory. In their classic book, *Perceptrons*, Minsky and Papert (1969) exploited primarily discrete mathematical methods that were compatible with the choice of binary units. They were incapable of analyzing any but the simplest learning networks. By changing the discrete threshold function of perceptrons to a smooth, differentiable curve, and thereby defining continuous-valued units, Rumelhart, Hinton, and Williams (1986) were able to apply continuous analytic methods to more complex learning networks. The result was a major advance in the power of subsymbolic learning.

A third historical example of the power of a continuous conception of subsymbolic computation relates to the connectionist generation of sequences. Traditionally this task has been viewed as making a connectionist system jump discretely between states to generate an arbitrary discrete sequence of actions $A_1, A_2, \cdots$. This view of the task reduces the connectionist system to a finite state machine that can offer little new to the analysis of sequential behavior. Recently Jordan (1986) has shown how a subsymbolic approach can give "for free" co-articulation effects where the manner in which actions are executed is influenced by future actions. Such effects are just what should come automatically from implementing serial behavior in a fundamentally parallel machine. Jordan's trick is to view the connectionist system as evolving continuously in time, with the task being the generation of a continuous trajectory through state space, a trajectory that meets as boundary conditions certain constraints, for example, that the discrete times 1, 2, $\cdots$ the system state must be in regions corresponding to the actions A_1, A_2, $\cdots$.

The final point is a foundational one. The theory of discrete computation is quite well understood. If there is any new theory of computation implicit in the subsymbolic approach, it is likely to be a result of a fundamentally different, continuous formulation of computation. It therefore seems fruitful, in order to maximize the opportunity for the subsymbolic paradigm to contribute new computational insights, to hypothesize that subsymbolic computation is fundamentally continuous.

It must be emphasized that the discrete/continuous distinction cannot be understood completely by looking at simulations. Discrete and continuous machines can of course simulate each other. The claim here is that the most analytically powerful descriptions of subsymbolic models are continuous ones, whereas those of symbolic models are not continuous.

This has profound significance because it means that many of the concepts used to understand cognition in the subsymbolic paradigm come from the category of continuous mathematics, while those used in the symbolic paradigm come nearly exclusively from discrete mathematics. Concepts from physics, from the theory of dynamical systems, are at least as likely to be important as concepts from the theory of digital computation. And analog computers, both electronic and optical, provide natural implementation media for subsymbolic systems (Anderson 1986; Cohen 1986).

8.2. Subsymbolic computation. An important illustration of the continuous/discrete mathematics contrast that distinguishes subsymbolic from symbolic computation is found in inference. A natural way to look at the knowledge stored in connections is to view each connection as a *soft constraint*. A positive (excitatory) connection from unit *a* to unit *b* represents a soft constraint to the effect that if *a* is active, then *b* should be too. A negative (inhibitory) connection represents the opposite constraint. The numerical magnitude of a connection represents the strength of the constraint.

Formalizing knowledge in soft constraints rather than hard rules has important consequences. Hard constraints have consequences singly; they are rules that can be applied separately and sequentially – the operation of each proceeding independently of whatever other rules may exist. But soft constraints have no implications singly; any one can be overridden by the others. It is only the entire set of soft constraints that has any implications. Inference must be a cooperative process, like the parallel relaxation processes typically found in subsymbolic systems. Furthermore, adding additional soft constraints can repeal conclusions that were formerly valid: Subsymbolic inference is fundamentally nonmonotonic.

One way of formalizing soft constraint satisfaction is in terms of statistical inference. In certain subsymbolic systems, the soft constraints can be identified as statistical parameters, and the activation passing procedures can be identified as statistical inference procedures (Geman & Geman 1984; Hinton & Sejnowski 1983b; Pearl 1985; Shastri 1985; Smolensky 1986a). This identification is usually rather complex and subtle: Unlike in classical "spreading activation" models and in many local connectionist models, the strength of the connection between two units is *not* determined solely by the correlation between their activity (or their "degree of association"). To implement subsymbolic statistical inference, the correct connection strength between two units will typically depend on all the other connection strengths. The subsymbolic learning procedures that sort out this interdependence through simple, strictly local, computations and ultimately assign the correct strength to each connection are performing no trivial task.

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

To sum up:

(25) a. Knowledge in subsymbolic computation is formalized as a large set of soft constraints.
b. Inference with soft constraints is fundamentally a parallel process.
c. Inference with soft constraints is fundamentally nonmonotonic.
d. Certain subsymbolic systems can be identified as using statistical inference.

9. Conceptual-level descriptions of intuition

The previous section concerned computation in subsymbolic systems analyzed at the subconceptual-level, the level of units and connections. In this final section I consider analyses of subsymbolic computation at the higher, conceptual level. Section 6 discussed subsymbolic modeling of conscious rule interpretation; here I consider subsymbolic models of intuitive processes. I will elaborate the point foreshadowed in Section 5: Conceptual-level descriptions of aspects of subsymbolic models of intuitive processing roughly approximate symbolic accounts. The picture that emerges is of a symbiosis between the symbolic and subsymbolic paradigms: The symbolic paradigm offers concepts for better understanding subsymbolic models, and those concepts are in turn illuminated with a fresh light by the subsymbolic paradigm.

9.1. The Best Fit Principle. The notion that each connection represents a soft constraint can be formulated at a higher level:

(26) The Best Fit Principle:
Given an input, a subsymbolic system outputs a set of inferences that, as a whole, gives a best fit to the input, in a statistical sense defined by the statistical knowledge stored in the system's connections.

In this vague form, this principle can be regarded as a desideratum of subsymbolic systems. Giving the principle, formal embodiment in a class of connectionist dynamical systems was the goal of harmony theory (Riley & Smolensky 1984; Smolensky 1983; 1984a; 1984b; 1986a; 1986c).

To render the Best Fit Principle precise, it is necessary to provide precise definitions of "inferences," "best fit," and "statistical knowledge stored in the system's connections." This is done in harmony theory, where the central object is the harmony function H which measures, for any possible set of inferences, the goodness of fit to the input with respect to the soft constraints stored in the connection strengths. The set of inferences with the largest value of H, that is, highest harmony, is the best set of inferences, with respect to a well-defined statistical problem.

Harmony theory offers three things. It gives a mathematically precise characterization of the prediction-from-examples goal as a statistical inference problem. It tells how the prediction goal can be achieved using a network with a certain set of connections. Moreover, it gives a procedure by which the network can learn the correct connections with experience, thereby satisfying the prediction-from-examples goal.

The units in harmony networks are stochastic: The differential equations defining the system are stochastic. There is a system parameter called the *computational temperature* that governs the degree of randomness in the units' behavior, it goes to zero as the computation proceeds. (The process is *simulated annealing*, like in the Boltzmann machine: Ackley, et al. 1985; Hinton & Sejnowski 1983a, 1983b, 1986. See Rumelhart, McClelland & the PDP Research Group, 1986, p. 148, and Smolensky, 1986a, for the relations between harmony theory and the Boltzmann machine.)

9.2. Productions, sequential processing, and logical inference. A simple harmony model of expert intuition in qualitative physics was described by Riley and Smolensky (1984) and Smolensky (1986a, 1986c). The model answers questions such as, "What happens to the voltages in this circuit if I increase this resistor?" (The questions refer to a particular simple circuit; the model's expertise is built in and not the result of learning.) This connectionist problem-solving system illustrates several points about the relations between subconceptual- and conceptual-level descriptions of subsymbolic computation.

Very briefly, the model looks like this. The state of the circuit is represented as a vector of activity over a set of network units we can call *circuit state feature units* – "feature units" for short. A subpart of this activity pattern represents whether the circuit's *current* has gone up, down, or stayed the same; other subparts indicate what has happened to the *voltage drops*, and so on. Some of these subpatterns are fixed by the givens in the problem, and the remainder comprise the answer to be computed by the network. There is a second set of network units, called *knowledge atoms*, each of which corresponds to a subpattern of activity over feature units. The subpatterns of features encoded by knowledge atoms are those that can appear in representations of possible states of the circuit: They are subpatterns that are allowed by the laws of circuit physics. The system's knowledge of Ohm's Law, for example, is distributed over the many knowledge atoms whose subpatterns encode the legal feature combinations for current, voltage, and resistance. The connections in the network determine which feature subpattern corresponds to a given knowledge atom. The subpattern corresponding to knowledge atom α includes a positive (negative) value for a particular feature f if there is a positive (negative) connection between unit α and unit f; the subpattern for α does not include f at all if there is no connection between α and f. All connections are two-way: Activity can propagate from feature units to knowledge atoms and vice versa. The soft constraints encoded by these connections, then, say that "if subpattern α is present, then feature f should be positive (negative), and vice versa."

In the course of computing an answer to a question, the units in the network change their values hundreds of times. Each time a unit recomputes its value, we have a *microdecision*. As the network converges to a solution, it is possible to identify *macrodecisions*, each of which amounts to a commitment of part of the network to a portion of the solution. Each macrodecision is the result of many individual microdecisions. These macrodecisions are approximately like the firing of production rules. In fact, these productions fire in essentially the same order

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

as in a symbolic forward-chaining inference system.[15] One can measure the total amount of order in the system and see that there is a qualitative change in the system when the first microdecisions are made – the system changes from a disordered phase to an ordered one.

It is a corollary of the way this network embodies the problem domain constraints, and the general theorems of harmony theory, that the system, when given a well-posed problem and unlimited relaxation time, will always give the correct answer. So under that idealization, the *competence* of the system is described by *hard* constraints: Ohm's Law, Kirchoff's Law – the laws of simple circuits. It's as though the model had those laws written down inside it. However, as in all subsymbolic systems, the *performance* of the system is achieved by satisfying a large set of *soft* constraints. What this means is that if we depart from the ideal conditions under which hard constraints seem to be obeyed, the illusion that the system has hard constraints inside is quickly dispelled. The system can violate Ohm's Law if it has to, but if it needn't violate the law, it won't. Outside the idealized domain of well-posed problems and unlimited processing time, the system gives sensible performance. It isn't brittle the way that symbolic inference systems are. If the system is given an ill-posed problem, it satisfies as many constraints as possible. If it is given inconsistent information, it doesn't fall flat and deduce just anything at all. If it is given insufficient information, it doesn't sit there and deduce nothing at all. Given limited processing time, the performance degrades gracefully as well. All these features emerge "for free," as automatic consequences of performing inference in a subsymbolic system; no extra machinery is added on to handle the deviations from ideal circumstances.

Returning to a physics level analogy introduced in Section 5, we have a "quantum" system that appears to be "Newtonian" under the proper conditions. A system that has, at the microlevel, soft constraints satisfied in parallel, has at the macrolevel, under the right circumstances, to have hard constraints, satisfied serially. But it doesn't *really*, and if you go outside the Newtonian domain, you see that it's really been a quantum system all along.

This model exemplifies the competence/performance distinction as it appears in the subsymbolic paradigm. We have an inference system (albeit a very limited one) whose performance is completely characterizable at the subconceptual-level in terms of standard subsymbolic computation: massively parallel satisfaction of multiple soft constraints. The system is fundamentally soft. Just the same, the behavior of the system can be analyzed at a higher level, and, under appropriate situations (well-posed problems), and under suitable processing idealizations (unlimited computation time), the competence of the system can be described in utterly different computational terms: The hard rules of the circuit domain. The competence theory is extremely important, but the performance theory uses radically different computational mechanisms.

The relation of the competence theory and the performance theory for this model can be viewed as follows. The behavior of the system is determined by its harmony function, which determines a surface or "landscape" of harmony values over the space of network states. In this landscape there are peaks where the harmony achieves its maximal value: These global maxima correspond to network states representing circuit conditions that satisfy all the laws of physics. The competence theory nicely describes the structure of this discrete constellation of global harmony maxima. But these maxima are a tiny subset of an extended harmony landscape in which they are embedded, and the network's performance is a stochastic search over the harmony landscape for these peaks. The givens of a problem restrict the search to the portion of the space consistent with those givens. If the problem is well-posed, exactly one of the global harmony peaks will be accessible to the system. Given unlimited search time, the system will provably end up at this peak: This is the limit in which the performance theory is governed by the competence theory. As the search time is reduced, the probability of the system's not ending up at the correct harmony peak increases. If insufficient information is given in the problem, multiple global harmony peaks will be accessible, and the system will converge to one of those peaks. If inconsistent information is given in the problem, none of the global harmony peaks will be accessible. But within the space of states accessible to the network there will be highest peaks of harmony – these peaks are not as high as the inaccessible global maxima; they correspond to network states representing circuit states that satisfy as many as possible of the circuit laws. As the network computes, it will converge toward these best-available peaks.

Subsymbolic computation is the evolution of a dynamical system. The input to the computation is a set of constraints on which states are accessible to the system (or, possibly, the state of the system at time zero). The dynamical system evolves in time under its defining differential equations; typically, it asymptotically approaches some equilibrium state – the output. The function relating the system's input to its output is its competence theory. This function is extremely important to characterize. But it is quite different from the performance theory of the system, which is the differential equation governing the system's moment-to-moment evolution. Relating the performance and competence of cognitive systems coincides with one of the principal tasks of dynamical systems theory: relating a system's local description (differential equations) to its global (asymptotic) behavior.

9.3. Conceptual-level spreading activation. In Section 7.2 it was pointed out that states of a subsymbolic model can be approximately analyzed as superpositions of vectors with individual conceptual-level semantics. It is possible to approximately analyze connectionist dynamical systems at the conceptual level, using the mathematics of the superposition operation. If the connectionist system is purely linear (so that the activity of each unit is precisely a weighted sum of the activities of the units giving it input), it can easily be proved that the higher-level description obeys formal laws of just the same sort as the lower level: The computations at the subconceptual and conceptual levels are isomorphic. Linear connectionist systems are of limited computational power, however; most interesting connectionist systems are nonlinear. Nevertheless, most of these are in fact *quasi*linear: A unit's value is computed by taking the weighted sum of its inputs and passing this

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

through a nonlinear function like a threshold or sigmoid. In quasi-linear systems, each unit combines its inputs linearly even though the effects of this combination on the unit's activity is nonlinear. Furthermore, the problem-specific knowledge in such systems is in the combination weights, that is, the linear part of the dynamical equations; and in learning systems, it is generally only these linear weights that adapt. For these reasons, even though the higher level is not isomorphic to the lower level in nonlinear systems, there are senses in which the higher level approximately obeys formal laws similar to the lower level. (For details, see Smolensky 1986b.)

The conclusion here is a rather different one from the preceding section, where we saw how there are senses in which higher-level characterizations of certain subsymbolic systems approximate productions, serial processing, and logical inference. Now what we see is that there are also senses in which the laws describing cognition at the conceptual level are activation-passing laws like those at the subconceptual-level but operating between units with individual conceptual semantics. Such semantic level descriptions of mental processing (which include *local* connectionist models; see note 3) have been of considerable value in cognitive science. We can now see how these "spreading activation" accounts of mental processing can fit into the subsymbolic paradigm.

9.4. Schemata. The final conceptual-level notion I will consider is that of the *schema* (e.g., Rumelhart 1980). This concept goes back at least to Kant (1787/1963) as a description of mental concepts and mental categories. Schemata appear in many AI systems in the forms of frames, scripts, or similar structures: They are prepackaged bundles of information that support inference in prototypical situations. [See also Arbib: "Levels of Modeling of Mechanisms of Visually Guided Behavior" *BBS* 10(3) 1987.]

Briefly, I will summarize work on schemata in connectionist systems reported in Rumelhart, Smolensky, McClelland & Hinton (1986) (see also Feldman 1981; Smolensky 1986a; 1986c). This work addressed the case of schemata for rooms. Subjects were asked to describe some imagined rooms using a set of 40 features like has-ceiling, has-window, contains-toilet, and so on. Statistics were computed on these data and were used to construct a network containing one node for each feature as well as connections computed from the statistical data.

The resulting network can perform inference of the same general kind as that carried out by symbolic systems with schemata for various types of rooms. The network is told that some room contains a ceiling and an oven; the question is, what else is likely to be in the room? The system settles down into a final state, and among the inferences contained in that final state are: the room contains a coffee cup but no fireplace, a coffee pot but no computer.

The inference process in this system is simply one of greedily maximizing harmony. [Cf. *BBS* multiple book review of Sperber & Wilson's *Relevance*, *BBS* 10(4).] To describe the inference of this system on a higher level, we can examine the global states of the system in terms of their harmony values. How internally consistent are the various states in the space? It's a 40-dimensional state space, but various 2-dimensional subspaces can be selected, and the harmony values there can be graphically displayed. The harmony landscape has various peaks; looking at the features of the state corresponding to one of the peaks, we find that it corresponds to a prototypical bathroom; others correspond to a prototypical office, and so on for all the kinds of rooms subjects were asked to describe. There are no *units* in this system for bathrooms or offices – there are just lower-level descriptors. The prototypical bathroom is a pattern of activation, and the system's recognition of its prototypicality is reflected in the harmony peak for that pattern. It is a consistent, "harmonious" combination of features: better than neighboring points, such as one representing a bathroom without a bathtub, which has distinctly lower harmony.

During inference, this system climbs directly uphill on the harmony landscape. When the system state is in the vicinity of the harmony peak representing the prototypical bathroom, the inferences it makes are governed by the shape of the harmony landscape there. This shape is like a schema that governs inferences about bathrooms. (In fact, harmony theory was created to give a connectionist formalization of the notion of schema; see Smolensky, 1984b; 1986a; 1986c.) Looking closely at the harmony landscape, we can see that the terrain around the "bathroom" peak has many of the properties of a bathroom schema: variables and constants, default values, schemata embedded inside schemata, and even cross-variable dependencies, which are rather difficult to incorporate into symbolic formalizations of schemata. The system behaves as though it had schemata for bathrooms, offices, and so forth, even though they are not really there at the fundamental level: These schemata are strictly properties of a higher-level description. They are informal, approximate descriptions – one might even say they are merely metaphorical descriptions – of an inference process too subtle to admit such high-level descriptions with great precision. Even though these schemata may not be the sort of object on which to base a formal model, nonetheless they are useful descriptions that help us understand a rather complex inference system.

9.5. Summary. In this section the symbolic structures in the intuitive processor have been viewed as entities in high-level descriptions of cognitive dynamical systems. From this perspective, these structures assume rather different forms from those arising in the symbolic paradigm. To sum up:

(27) a. Macroinference is not a process of firing a symbolic production but rather of qualitative state change in a dynamical system, such as a phase transition.

b. Schemata are not large symbolic data structures but rather the potentially intricate shapes of harmony maxima.

c. Categories (it turns out) are attractors in connectionist dynamical systems: states that "suck in" to a common place many nearby states, like peaks of harmony functions.

d. Categorization is not the execution of a symbolic algorithm but rather the continuous evolution of the dynamical system – the evolution that drives states into the attractors that maximize harmony.

e. Learning is not the construction and editing of formulae, but rather the gradual adjustment of connection strengths with experience, with the effect of

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

slowly shifting harmony landscapes, adapting old and creating new concepts, categories, and schemata.

The heterogeneous assortment of high-level mental structures that have been embraced in this section suggests that the conceptual-level lacks formal unity. This is precisely what one expects of approximate higher-level descriptions, which, capturing different aspects of global properties, can have quite different characters. According to the subsymbolic paradigm, the unity underlying cognition is to be found not at the conceptual level, but rather at the subconceptual level, where relatively few principles in a single formal framework lead to a rich variety of global behaviors.

10. Conclusion

In this target article I have not argued for the validity of a connectionist approach to cognitive modeling, but rather for a particular view of the role a connectionist approach might play in cognitive science. An important question remains: Should the goal of connectionist research be to replace other methodologies in cognitive science? Here it is important to avoid the confusion discussed in Section 2.1. There I argued that for the purpose of science, it is sound to formalize knowledge in linguistically expressed laws and rules – but it does not follow therefore that knowledge in an individual's mind is best formalized by such rules. It is equally true that even if the knowledge in a native speaker's mind is well formalized by a huge mass of connection strengths, it does not follow that the science of language should be such a set of numbers. On the contrary, the argument of Section 2.1 implies that the science of language should be a set of linguistically expressed laws, to the maximal extent possible.

The view that the goal of connectionist research should be to replace other methodologies may represent a naive form of eliminative reductionism. Successful lower-level theories generally serve not to replace higher-level ones, but to enrich them, to explain their successes and failures, to fill in where the higher-level theories are inadequate, and to unify disparate higher-level accounts. The goal of subsymbolic research should not be to replace symbolic cognitive science, but rather to explain the strengths and weaknesses of existing symbolic theory, to explain how symbolic computation can emerge out of nonsymbolic computation, to enrich conceptual-level research with new computational concepts and techniques that reflect an understanding of how conceptual-level theoretical constructs emerge from subconceptual computation, to provide a uniform subconceptual theory from which the multiplicity of conceptual theories can all be seen to emerge, to develop new empirical methodologies that reveal subconceptual regularities of cognitive behavior that are invisible at the conceptual level, and to provide new subconceptual-level cognitive principles that explain these regularities.

The rich behavior displayed by cognitive systems has the paradoxical character of appearing on the one hand tightly governed by complex systems of hard rules, and on the other to be awash with variance, deviation, exception, and a degree of flexibility and fluidity that has quite eluded our attempts at simulation. *Homo sapiens* is the rational animal, with a mental life ruled by the hard laws of logic – but real human behavior is riddled with strong nonrational tendencies that display a systematicity of their own. Human language is an intricate crystal defined by tight sets of intertwining constraints – but real linguistic behavior is remarkably robust under deviations from those constraints. This ancient paradox has produced a deep chasm in both the philosophy and the science of mind: on one side, those placing the essence of intelligent behavior in the hardness of mental competence; on the other, those placing it in the subtle softness of human performance.

The subsymbolic paradigm suggests a solution to this paradox. It provides a formal framework for studying how a cognitive system can possess knowledge which is fundamentally *soft*, but at the same time, under ideal circumstances, admit good higher-level descriptions that are undeniably *hard*. The passage from the lower, subconceptual level of analysis to the higher, conceptual level naturally and inevitably introduces changes in the character of the subsymbolic system: The computation that emerges at the higher level incorporates elements with a nature profoundly different from that of the fundamental computational processes.

To turn this story into a scientific reality, a multitude of serious conceptual and technical obstacles must be overcome. The story does, however, seem to merit serious consideration. It is to be hoped that the story's appeal will prove sufficient to sustain the intense effort that will be required to tackle the obstacles.

ACKNOWLEDGMENTS

I am indebted to Dave Rumelhart for several years of provocative conversations on many of these issues; his contributions permeate the ideas formulated here. Sincere thanks to Jerry Fodor and Zenon Pylyshyn for most instructive conversations. Comments on earlier drafts from Geoff Hinton, Mark Fanty, and Dan Lloyd were very helpful, as were pointers from Kathleen Akins. Extended comments on the manuscript by Georges Rey were extremely helpful. I am particularly grateful for a number of insights that Rob Cummins and Denise Dellarosa have generously contributed to this paper.

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.

NOTES

1. In this target article, when *interpretation* is used to refer to a process, the sense intended is that of computer science: the process of taking a linguistic description of a procedure and executing that procedure.

2. Consider, for example, the connectionist symposium at the University of Geneva held Sept. 9, 1986. The advertised program featured Feldman, Minsky, Rumelhart, Sejnowski, and Waltz. Of these five researchers, three were major contributors to the symbolic paradigm for many years (Minsky 1975; Rumelhart 1975; 1980; Waltz 1978).

3. This is an issue that divides connectionist approaches. "Local connectionist models" (e.g., Dell 1985; Feldman 1985; McClelland & Rumelhart 1981; Rumelhart & McClelland 1982; Waltz & Pollack 1985) accept (9), and often deviate significantly from (8a). This approach has been championed by the Rochester connectionists (Feldman et al. 1985). Like the symbolic paradigm, this school favors simple semantics and more complex operations. The processors in their networks are usually more powerful than those allowed by (8); they are often like digital computers running a few lines of simple code. ("If there is a 1 on this input line then do X else do Y," where X and Y are quite

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

different simple procedures; e.g., Shastri 1985.) This style of connectionism, quite different from the subsymbolic style, has much in common with techniques of traditional computer science for "parallelizing" serial algorithms by decomposing them into routines that can run in parallel, often with certain synchronization points built in. The grain size of the Rochester parallelism, although large compared to the subsymbolic paradigm, is small compared to standard parallel programming: The processors are allowed only a few internal states and can transmit only a few different values (Feldman & Ballard 1982).

4. As indicated in the introduction, a sizeable sample of research that by and large falls under the subsymbolic paradigm can be found in the books, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition:* Rumelhart, McClelland, and the PDP Research Group 1986; McClelland, Rumelhart, and the PDP Research Group 1986. While this work has since come to be labelled "connectionist," the term "PDP" was deliberately chosen to distinguish it from the localist approach, which had previously adopted the name "connectionist" (Feldman & Ballard 1982).

5. The phrase is Roger Schank's, in reference to "parallel processing" (Waldrop 1984). Whether he was referring to connectionist systems I do not know; in any event, I don't mean to imply that the grounds for his comment are addressed here.

6. In this section the disclaimer in the introduction is particularly relevant: The arguments I offer are not intended to represent a consensus among connectionists.

7. For example, two recently discovered learning rules that allow the training of hidden units, the Boltzmann machine learning procedure (Hinton & Sejnowski 1983a) and the back-propagation procedure (Rumelhart, Hinton & Williams 1986), both involve introducing computational machinery that is motivated purely mathematically; the neural counterparts of which are so far unknown (unit-by-unit connection strength symmetry, alternating Hebbian and antiHebbian learning, simulated annealing, and backwards error propagation along connections of identical strength to forward activation propagation).

8. A notable exception is Touretzky and Hinton 1985.

9. Furthermore, when a network makes a mistake, it can be told the correct answer, but it cannot be told the precise rule it violated. Thus it must assign blame for its error in an undirected way. It is quite plausible that the large amount of training currently required by subsymbolic systems could be significantly reduced if blame could be focused by citing violated rules.

10. There is a trade-off between the number of goal conditions one chooses to attribute to a system, and the corresponding range of tolerable environmental conditions. Considering a large variety of environmental conditions for a river, there is only the "flow downhill" goal; by appropriately narrowing the class of conditions, one can increase the corresponding goal repertoire. A river can meet the goal of carrying messages from A to B, if A and B are appropriately restricted. But a homing pigeon can meet this goal over a much greater variety of situations.

11. Consider a model that physicists like to apply to "neural nets" – the *spin glass* (Toulouse et al. 1986). Spin glasses seem relevant because they are dynamical systems in which the interactions of the variables ("spins") are spatially inhomogeneous. But a spin glass is a system in which the interactions between spins are random variables that all obey the *same* probability distribution p: The system has *homogeneous inhomogeneity*. The analysis of spin glasses relates the properties of p to the bulk properties of the medium as a whole; the analysis of a single spin subject to a particular set of inhomogeneous interactions is regarded as quite meaningless, and techniques for such analysis are not generally developed.

12. This problem is closely related to the localization of a failure of veridicality in a scientific theory. Pursuing the remarks of Section 2.1, scientific theories can be viewed as cognitive systems, indeed ones having the prediction goal. Veridicality is a property of a scientific theory as a whole, gauged ultimately by the success or failure of the theory to meet the prediction goal. The veridicality of abstract representations in a theory derives solely from their causal role in the accurate predictions of observable representations.

13. For example, in both harmony theory and the Boltzmann machine, discrete units have typically been used because (1) discrete units simplify both analysis and simulation; (2) for the quadratic harmony or energy functions that are being optimized, it can be proved that no optima are lost by simplifying to binary values; (3) these models' stochastic search has a "jumpy" quality to it anyway. These, at least, are the *computational* reasons for discrete units; in the case of the Boltzmann machine, the discrete nature of action potentials is also cited as a motivation for discrete units (Hinton et al. 1984).

14. Alternatively, if the original harmony/Boltzmann approach is extended to include nonquadratic harmony/energy functions, nonbinary optima appear, so again one switches to continuous units (Derthick, in progress; Smolensky, in progress).

15. Note that these (procedural) "productions" that occur in intuitive processing are very different from the (declarative) production rules of Section 6 that occur in conscious rule application.

Open Peer Commentary

Commentaries submitted by the qualified professional readership of this journal will be considered for publication in a later issue as Continuing Commentary on this article. Integrative overviews and syntheses are especially encouraged.

On the proper treatment of the connection between connectionism and symbolism

Louise Antony and Joseph Levine

Department of Philosophy & Religion, North Carolina State University, Raleigh, N. C. 27695

Smolensky is concerned to establish two claims: first, that there is a genuine conflict between the connectionist and the classical computationalist (symbolic) approaches to the study of mind; and second, that the conflict consists in a disagreement about the level of analysis at which it is possible to obtain an accurate and complete account of cognitive processing. These two claims, we contend, are incompatible. If the difference between connectionism and symbolism really concerns levels of theorizing, then there is no incompatibility. If there is a genuine conflict, then talk of alternative levels or degrees of approximation is either misleading or wrong.

In Smolensky's picture, connectionism and symbolism share a problem domain: They both seek to explain cognitive processes. Moreover, they both accept an initial analysis of the problem domain effected at what Smolensky calls the conceptual level. At this level, cognitive processes are characterized as the operation of rules defined over representations. This level provides a good approximation of conscious cognitive processes, and a passable account of intuitive processes. It is on the issue of how to provide more precise accounts of intuitive processing that the two paradigms diverge: The symbolists posit the same

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

kinds of entities and mechanisms described at the conceptual level; the subsymbolists shift domains downward, to the subconceptual level, where neither the primitive entities nor the basic operations map onto anything at the conceptual level.

But the nature of this downshift is obscure, and attempts to clarify it make problems for one or the other of Smolensky's central claims. One suggested mark of a domain shift is semantic: a shift from the consciously accessible concepts encoded in natural language to finer-grained microfeatures.[1] But this won't do; many paradigmatically symbolic cognitive theories propose computational processes defined over elements of both sorts. Indeed, the positing of unconscious manipulations of elements with semantics *unavailable* to conscious introspection is part and parcel of most symbolic cognitive accounts. Decompositional semantics, for example, hypothesized that the semantic primitives of natural language (horse, kill) were subconsciously decomposed into a set of conceptual primitives, which included things such as categorized variables, or markers like +inchoative, that few native speakers would own. Phonology, which constructs morphemes out of phonemes and phonemes out of phones, provides another example. Any such theory may be wrong (compositional semantics almost certainly is) but each is symbolic par excellence.

The mark of a domain shift, therefore, cannot be purely semantic. The other possibility is that a domain shift involves a change in the kinds of mechanisms posited. This criterion reveals a deep difference between the two paradigms. The central mechanism of the symbolic paradigm is the structure-sensitive operation. Such mechanisms are posited at every genuinely cognitive level of theorizing. In the connectionist paradigm, however, structure-sensitive operations posited at the conceptual level provide only a rough approximation of actual cognitive processing. A full and accurate account can be found only at the subconceptual level, where symbolic processes are replaced by numerical functions describing state changes in a dynamical system.

We know that the theory of the ultimate physical implementation of any symbolic process will certainly need to hypothesize nonsyntactic mechanisms. But Smolensky maintains that connectionism should not be regarded as simply an implementation theory for symbolic paradigm theories. Connectionist theories, unlike theories of implementation, are cognitive. Connectionism and symbolism offer two competing models of the same cognitive capacities, models that posit strikingly different mechanisms to explain the input–output functions definitive of those capacities.

The relation between the two paradigms, Smolensky argues, is analogous to the relation between classical and quantum mechanics: The symbolic model gives an approximately true description of the goings-on precisely characterized by a mathematical description of the dynamical systems that actually run intuitive cognitive processes. But there's the rub. Given the radical difference between the mechanisms and mode of explanation posited by the symbolic and the connectionist paradigms, what could it mean to say that, from a connectionist perspective, a symbolic theory is even "approximately" true?

The symbolic paradigm claims to explain cognitive phenomena by hypothesizing the real existence of structured representations, and operations defined over them. One thing it could mean to say – that a symbolic model is only approximately correct – is that the model is an idealization, an abstract characterization of a system, which, once physically realized, is subject to glitches deriving from the physical nature of the realizing medium. (Economic theories, for example, usually assume higher quality hardware than is generally available.) Alternatively, one could mean that the model describes only one subsystem of a complex system, and thus cannot be used to predict precisely the behavior of the whole. (This is our understanding – contrary to Smolensky's – of the relation Chomsky posits between linguistic competence and linguistic performance.) [See Chomsky: "Rules and Representations" *BBS* 3(1) 1980.]

What one shouldn't mean is simply that the model gets the input–output relations right. But if Smolensky is right, if connectionist models are not models of the implementation of symbolic processes, that's all it could mean to say that symbolic theories are approximately true; for there are no structure-sensitive operations in reality, and there are no structured constituents. Cognitive models are only correct (and then only approximately so) in their mappings of inputs to outputs. Their explanations of why those mappings hold are fantasy.

Smolensky's allowance that a sort of constituent structure can be read into connectionist networks makes no difference. The crucial question is whether that structure is implicated in the explanation of cognitive processes or not. If it is not, then the appearance of constituent structure is accidental and unexplained, as indeed it ought to be given the nature of connectionist mechanisms.[2] If, on the other hand, constituent structure is built into the operations of the network – if, say, the initial connection strengths and learning functions serve to ensure that constituency relations are respected in state transitions, then the organization of the networks is constrained by the symbolic processes posited at the conceptual level. In that case, the network would be simply an implementation of the symbolic processes.

In sum: If the connectionist and symbolic paradigms are indeed incompatible, it is because at the same explanatory level they disagree fundamentally about the nature of mental processes. If, on the other hand, both models yield explanatory insight into the workings of the mind, though at different levels of description, then we must understand connectionist models as implementation models.[3]

NOTES

1. This criterion is suggested by Smolensky's remarks in sect. 3, para. 2; and also by the discussion of reduction and instantation in sect. 5, para. 6.

2. Fodor and Pylyshyn develop this point in detail in "Connectionism and Cognitive Architecture: A Critical Analysis" (unpublished manuscript). Smolensky explicitly addresses their arguments in sect. 7.1., but, in our view, does not answer them adequately.

3. We would like to thank David Auerbach and Harold Levin for helpful discussions of this paper.

Connectionism and interlevel relations

William Bechtel

Department of Philosophy, Georgia State University, Atlanta, Ga. 30303

Smolensky's proposal to treat connectionist models as applying to an intermediate ("subconceptual" level) between neural models and conceptual models is a very attractive one. I am troubled, though, by the way he articulates the relation among the three levels. There seem to be two different ways these levels relate: The higher level may simply provide a more abstract characterization of the lower level, or it may actually constitute a higher level in a part–whole hierarchy. There are important differences between these types of relationships. To begin, one is a relationship between theories, whereas the other represents an ontological distinction in nature which has consequences for theorizing. I suspect that for Smolensky the relationship between the neural and the subconceptual level is of the first sort, and that between the subconceptual and the conceptual level is of the second sort. These types of relationships, though, have different consequences.

Consider first the relation between neural and subconceptual levels. Smolensky characterizes the connectionist models at the subconceptual level as syntactically more like neural models but semantically more like symbolic models. The reason they are semantically more like symbolic models is that current neural

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

theories lack an adequate semantic account. Below I shall claim that any semantic account developed for the neural level will also be similar to that of symbolic accounts. The critical issues for level relations arise on the syntactic side. There are a number of differences, Smolensky notes, between the syntax of neural processing models and that of connectionist models, and he even comments that currently "architectural decisions [for connectionist systems] seem to be driven more and more by mathematical considerations and less and less by neural ones" (sect. 4, para. 9). These differences could support the claim that we are dealing with different levels of organization in nature. But this raises the question: What do connectionist systems model if not the activity of neuronal networks? There may be units in the brain at a level above that of neural networks which fit the characteristics of the connectionist architecture better than neural networks, but I do not see any reason to expect this result. Moreover, in his discussion of how models like NETtalk can inform us of general characteristics of a wide class of systems, Smolensky gives us reason to think that connectionist accounts are simply more abstract and currently more tractable than neural accounts. They may provide us with reasonably accurate approximations of the actual performance of neural networks, which will enable us to carry on with theorizing about mental phenomena in the absence of detailed neural models. Although I do not find this problematic, it does mean that connectionist and neural theories (when developed in order to understand cognitive performance) are characterizing the same phenomena in nature. It may prove useful at a later stage in the inquiry to amend connectionist models to better accommodate knowledge of the nervous system, and then the distinction between neural and cognitive models may largely disappear.

Now consider the relationship between connectionist and symbolic models. Here we seem to be dealing with a genuine part–whole relation. The interactions between the units in the connectionist model give rise to roughly stable patterns, which are then assigned the semantics of conscious concepts (see thesis 8b). Smolensky discusses the relationships between connectionist and conceptual models in the same way he discussed the relationship between neural and connectionist accounts – the accounts are approximately equivalent. Moreover, for him it is important that the agreement is only approximate and that symbolic processes are not directly implementable in the subconceptual system (thesis 8c). What does this say about the status of the conceptual level?

There are, as Smolensky notes at the beginning of the target article, a variety of human endeavors (particularly interpersonal activities, such as scientific investigations) in which symbols and conscious symbol manipulation are important.[1] If we take a realist view of the symbols and symbol processing that occur in these activities, we should expect them to be implemented in lower level activity: These symbols and symbol processing activities should be the causal product of the processes at the lower level and thus implemented by the lower level processes. But thesis 8c says that these symbolic processes are not directly implementable, so how do they occur? Are they real processes in nature? It is this which seems to make connectionism a version of eliminative materialism.

There is a way to adopt both this realistic view of the relation of explanatory accounts at different levels and Smolensky's thesis 8c. In most genuine interlevel relationships in science, study at one level tends to guide revision in theorizing at the other level. If this were to happen in the current case, we should expect work on the subconceptual level to lead to revisions in our understanding of the conceptual level. Thesis 8c should then be viewed as applying to current symbolic accounts, not to the accounts revised in light of work at the subconceptual level. One example of the changes that might result is a revision in the view of the concepts that figure in conscious thinking as fixed, stable units that are stored in memory and retrieved into working memory where they are manipulated by rules.

Connectionism may lead us to view concepts as far more temporary patterns, which change over time as a result of learning or other activity in the system. (Barsalou, 1986 and in press, has already produced evidence of a variability in concepts that would fit such an analysis.) Thus, if the relation of the subconceptual to the symbolic is really an interlevel one, we should not treat one account as simply a rough approximation of the other, to be surrendered if we develop a better approximation, but we should view the accounts as descriptions of different phenomena related in a part–whole manner, with the goal being a causal explanation of the higher level phenomena in terms of the lower (Bechtel 1988).

In closing I wish to return to something I noted above: One reason connectionist accounts are very like conceptual-level accounts semantically is that any semantic account (even one at the neural level) will be quite similar to that at the conceptual level. We can see this by considering Smolensky's brief remarks about the semantics of the subconceptual level. Our semantic interpretations of processes in a system at any level depend on how those processes enable the system to meet its environmental goals. Interpretations of this sort can provide the intentional perspective for the subconceptual level that Smolensky earlier suggests is lacking for connectionist systems. To capture what is often viewed as a defining feature of intentional systems – the ability of their states to be about nonexisting phenomena – we need to bear in mind that no such system is perfectly adapted to its environment any more than any organism is; our intentional interpretation of its states will also have to show how it fails to satisfy its environmental goals (Bechtel 1985).

NOTE

1. My concern is only with conscious rule processing, not with intuitive processing. The study of the latter, for which rule-processing accounts have been generally inadequate, should perhaps be transferred to the subconceptual level.

Two constructive themes

Richard K. Belew

Computer Science and Engineering Department, University of California at San Diego, La Jolla, Calif. 92093

Connectionism is definitely hot these days, and one of the obvious questions is how these models relate to previous ones. With this target article, Smolensky clearly pushes the debate to the next plateau. He argues that there is an important, valid level of cognitive modeling below the conventional symbolic level yet above the level of neural modeling. If his paper has a flaw, it is that Smolensky spends more time rhetorically delineating how his subsymbolic models are not either symbolic or neural, rather than constructively emphasizing those characteristics that make subsymbolic models important and valid. He does mention some of these characteristics as subthemes, however, and I think it is worth emphasizing two in particular.

One of Smolensky's first distinctions is between cultural and individual knowledge. Individual knowledge, that information used by a single person to help him function in the world, has typically been the provenance of cognitive science. Smolensky identifies a second type of cultural knowledge that groups of people collectively codify, share, learn, and use. He also makes the strong claim (his point 2b) that "We can view the top-level conscious processor of individual people as a virtual machine – the conscious rule interpreter – and we can view cultural knowledge as a program running on that machine" (sect. 2.1., para. 4). Smolensky then argues that the symbolic tradition in cognitive science – because of its preoccupation with cognitive processes of which we are consciously aware – has come to model the form of cultural rather than individual knowledge. After noting that "the constraints on cultural knowledge for-

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

malization are not the same as those on individual knowledge formalization," Smolensky then infers that a large body of subconscious cognitive phenomena is ripe for subsymbolic modeling.

Two observations about this argument should be made. First, we can accept the importance of Smolensky's distinction between cultural and individual knowledge without accepting this particular view of their relationship. I too have argued for the importance of cultural knowledge to cognitive science (Belew 1986, Ch. 8), but although Smolensky's hypothesis regarding the conscious rule interpreter is intriguing it is currently also unsubstantiated and, I fear, oversimplified.

Second, Smolensky fails to draw an important conclusion from this argument, namely, that subsymbolic models have a distinct disadvantage to symbolic ones in that they cannot easily be assimilated into the cultural knowledge acquisition process known as science. There is bound to be a bias on the part of unwary cognitive scientists toward "linguistic," symbolic models simply because these are more easily communicated from one scientist to another and hence easily considered more "scientific."

Subsymbolic models rest on the assumption that some of the most interesting cognitive phenomena cannot be modeled in terms of symbol manipulation (Smolensky presents this as point 7c); that is, the rejection of Newell's Physical System Hypothesis (1980) (which Smolensky effectively restates as his strawman point 3a). These models must therefore be stated in nonlinguistic terms, typically mathematical analyses, that are much more difficult for most cognitive scientists to appreciate. Such a bias is understandable, but if we want at least to allow for the possibility of subsymbolic models that cannot be expressed in easily understood symbols, we must be willing to dig in and understand the mathematics.

Luckily, one of the reasons for the current connectionist renaissance is that Smolensky and others have found new ways to make their mathematics more comprehensible. In particular, a key characteristic of these models is that they explore ways in which the dynamic rather than the structural characteristics of a system can contribute to intelligent behavior. This is a second critically important subtheme of Smolensky's analysis.

It seems quite obvious that AI and cognitive science would not be here if not for the computer. More insidious is the way our models of cognition have suffered as a result of the "von Neumann bottleneck" of sequential computation, imposed until very recently by existing computers. Of course, Newell and others have argued forcibly for the need to describe cognition as a basically sequential process (primarily because all cognitive systems must act in the sequential flow of time), but the correspondence between their sequential simulations and the availability of only sequential von Neumann computers is a bit too neat to be coincidental.

One of the most striking features of sequential models is their particularly simple dynamics: One thing happens at a time, and at one place. Using only this spare dynamic model, cognitive models and AI knowledge representations have focused on building elaborate structural systems. We have come to believe that building more intelligent systems means designing their knowledge structures more effectively.

One of connectionism's most distinctive features is that it views the dynamics of the cognitive system as just as important as its structure. In fact, by current knowledge representation standards, connectionist nets are particularly simple: weighted digraphs. How could this simple representation support intelligent behavior when we still have problems getting our most sophisticated semantic networks to work! The connectionist answer is that the current division between structure and dynamics is inappropriate. Much of our cognitive activity can and should be described in dynamical terms (the spreading of activation, the modulation of activity, etc.) and our models must be dynamically sophisticated as well.

I believe it is no coincidence that two of the most outspoken proponents of subsymbolic models, Smolensky and Hofstadter, were trained originally in physics, a science in which the structure and dynamics of systems are inextricably linked. A major contribution of these scientists has been to bring, from physics, the descriptive language of dynamical systems. This has begun to allow us to constructively model (and describe to other scientists) dynamical properties of our models that would otherwise be beyond our grasp.

It is important not to take Smolensky and connectionism too literally. Smolensky's "connectionist dynamical system" hypothesis (his point 7a) assumes a particular structural representation, a weighted digraph with a vector of activities, but many of the strengths and weaknesses of this computational system are exhibited by other representations, for example Holland's Classifier System (Holland et al. 1986). The Classifier System uses a representation, derived from production rules, that cannot be called connectionist but that is properly considered subsymbolic. A careful comparison of these two representations would not be appropriate here; I simply note that both representations embrace complex dynamic processing as an integral part of their operation. From the perspective of these newer representations, a key issue separating sequential, symbolic accounts from dynamical, subsymbolic ones is how sequential action emerges from parallel cogitation.

Smolensky's target article obviously does not answer this ambitious question; his purpose is simply "to formulate rather than argue the scientific merits of a connectionist approach." He has succeeded ably at outlining some basic tenets of subsymbolic cognitive modeling, and the stage is now set for further debate.

Information processing abstractions: The message still counts more than the medium

B. Chandrasekaran, Ashok Goel, and Dean Allemang

Laboratory for Artificial Intelligence Research, Department of Computer and Information Science, Ohio State University, Columbus, Ohio 43210

Smolensky's target article has two major virtues. First, it is a very clear presentation of the essential commitments of connectionism in the subsymbolic paradigm. Second, his claims on its behalf are relatively modest: He identifies one level, viz. the subconceptual level, as the appropriate province for connectionism, while leaving other levels as the domains for other kinds of theories. We find ourselves in agreement with Smolensky on several counts:

A satisfactory account of cognitive phenomena has to be representational.

Subsymbolic models are not merely implementations of symbolic models just because continuous functions can be simulated by Turing machines. As one of us argues in Chandrasekaran (1988), connectionist and symbolic methods of computing a function may make significantly different representational commitments about what is being represented, and thus may constitute different theories about an information process.

Finally, theories which use only conceptual entities accessible to the conscious level are likely to be inadequate to cover the range of phenomena in cognition. In fact, we regard much of the work on knowledge representation in the logic paradigm, where "thinking" is closely associated with the phenomena and mechanisms of conscious reasoning, as suffering from this problem.

However, Smolensky is not making enough of a distinction between what is being computed and the mechanisms of that computation. It is true that connectionism offers a medium of representation and mechanisms of processing different from those of the traditional symbolic paradigm. We believe, however, that computational leverage is more in the content of repre-

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

sentation than in the representational medium or processing mechanism. In this we follow Marr's (1982) proposal for an information processing (IP) level description of a theory, that specifies three kinds of information: (1) what is available at the input to a process, (2) what is needed at the output, and (3) what are the kinds of information that need to be made available as part of the process. The content of the theory is this set of information processing abstractions by means of which the input can be transformed to the output. Commitments about *how* these abstractions are represented and processed are made at the next level, where a number of symbolic or connectionist alternatives may be proposed. In contrast to Smolensky's proposal about how to carve up the competence/performance distinction, we suggest that the competence is represented by the IP abstraction level, and connectionist networks or symbolic algorithms are alternate realizations of this theory, leading to possible differences in the details of performance.

Viewed in this light, what is significant about the two tense-learning examples contrasted by Smolensky is that they propose different sets of IP abstractions: One set seems too close to the conscious level to be quite right, while the other proposes abstractions (e.g., "rounded") that are genuinely theoretical constructs and not consciously accessible. Good theories of complex phenomena are apt to involve primitives which are not accessible to naive consciousness. This is no less true of cognition than it is of physics. If the connectionist theory of learning of tense-endings has good performance, it is due as much to the particular IP abstractions that are represented in it as it is due to the medium of representation and the mechanisms. The issue of whether the primitive objects are close to consciousness is orthogonal to whether they are represented and manipulated connectionistically or algorithmically. Smolensky's conflation of these two issues is a special case of the general instance, common to all stripes of AI, of ascribing credit to mechanisms when quite a bit of it has to go to the IP theory that is being realized by the mechanisms. For instance, the content contributions of many nominally symbolic theories in AI are really at the level of IP abstractions to which they make a commitment, and not at the level of their implementation in a symbol structure. Symbols have often merely stood in for the abstractions that need to be captured one way or another. Note that we are not claiming that the medium of representation and manipulation does not make for important differences, but that attribution of the differences to the medium requires first an analysis of the role played by the IP abstractions.

Of course, if connectionism can provide learning mechanisms such that an agent can start out with few such abstractions and can learn to perform the IP function in a reasonable amount of time, then connectionism can sidestep most of the representational problems. The fundamental problem of complex learning is the credit (or blame) assignment problem, which Smolensky admits is very hard, but then somewhat startlingly claims has been largely solved by connectionism. However, if one looks at particular connectionist schemes that have been proposed for learning tasks, a significant part of the IP abstractions are built into the architecture in the choice of input, feedback directions, allocation of subnetworks, and the semantics that underlie the choice of layers, and so on. The inputs and the initial configuration incorporate a sufficiently large part of the requisite IP abstractions which constrain the search space, so that what is left to be learned, while nontrivial, is proportionately small. In fact, the search space is small enough so that statistical associations, for which connectionist learning mechanisms are particularly suited, can do the trick. This is not to downplay the contributions of the various propagation schemes, but to emphasize the role of the IP abstractions implicit in the networks even before the learning mechanisms begin to operate. In short, while connectionist mechanisms may be able to explain how learning can be accomplished as a series of searches in appropriate parameter spaces, they do not absolve the theorist of the responsibility to provide sufficient content to the theory in the form of a priori commitments made by the architecture.

Smolensky's conscious/intuitive and symbolic/connectionist distinctions again are orthogonal. Theory making with entities at the conscious level alone is not a problem of symbolic theories per se. For example, Schank's (1972) Conceptual Dependency theories and our own work on generic tasks in problem solving in Chandrasekaran (1987), both of which do not correspond to the terms in conscious reasoning. A major task of cognitive theory making is finding the right set of primitive terms, conscious, intuitive, or otherwise, that need to be represented. This task doesn't change, whether or not one's approach is connectionist.

We regard connectionism as an important corrective to the extreme view of the cognitive processor as nothing but a Turing machine. Connectionism offers intriguing insights into how some objects in symbolic theories, such as frames and schemas, may be composed at "run-time" from more diffuse connectionist representations. Our perspective on how the symbolic paradigm and connectionism coexist is a little different from that of Smolensky. Connectionist and symbolic computationalist phenomena have different but overlapping domains. Connectionist architectures seem to be especially good in providing some basic functions, such as retrieval, matching and low-level parameter learning, with intuitively desirable "micro" performance features such as speed and softness. Symbolic cognitive theories can take advantage of the availability of connectionist realizations of these functions in order to achieve greater fidelity in their modeling. Even here, the content theories have to be done just right for the mechanisms to work. For example, in retrieval, the basic problem will remain encoding of objects in memory with appropriate indices, except that now the representation of these encodings and the retrieval of the relevant objects may be done in the connectionist framework.

Much of cognitive theory making will and should remain largely unaffected by connectionism. We have given two reasons for this. First, most of the work is in coming up with an information processing theory in the first place. Second, none of the connectionist arguments or empirical results have shown that the symbolic, algorithmic character of a significant part of high level thought, at least in the macro level, is either a mistaken hypothesis, purely epiphenomenal, or simply irrelevant.

ACKNOWLEDGMENT

We acknowledge the support of the U.S. Air Force Office of Scientific Research (AFOSR-87-0090), and the U.S. Defense Advanced Research Projects Agency (RADC-F30602-85-C-0010).

Is Smolensky's treatment of connectionism on the level?

Carol E. Cleland

Department of Philosophy and Institute of Cognitive Science, University of Colorado, Boulder, Colo. 80309

In his very interesting target article, Smolensky remarks that "most of the foundational issues surrounding the connectionist approach turn, in one way or another, on the *level of analysis* adopted" (sect. 1.3., para. 1). From a philosophical point of view, one of the main novelties of Smolensky's PTC (proper treatment of connectionism) approach is the introduction of a new level for the analysis of cognition; in addition to the traditional conceptual and neural levels, there is the subconceptual level. Smolensky's claim is that the complete formal account of cognition lies at this (the subconceptual) level.

In exactly what sense is the subconceptual level supposed to be the more fundamental level for the analysis of cognition?

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

Smolensky explicitly likens the relationship between symbolic accounts of cognition (which adopt the conceptual level of analysis) and subsymbolic accounts of cognition (which adopt the subconceptual level of analysis) to the relationship between macrophysical accounts of physical phenomena and microphysical accounts of physical phenomena, for example, the relationship of classical mechanics to quantum mechanics, the relationship of classical thermodynamics to statistical thermodynamics (sect. 5, para. 11). The idea is that just as classical mechanics accurately describes the macrostructure of physical reality and quantum mechanics accurately describes the microstructure of physical reality, so symbolic models accurately describe the macrostructure of cognition and subsymbolic models accurately describe the microstructure of cognition. He concludes that symbolic accounts are "reducible" to subsymbolic accounts in the same sense that microphysics is "reducible" to macrophysics. This, then, appears to be the sense in which the subconceptual level is supposed to be more fundamental than the conceptual level: Just as rocks and chairs are nothing more than collections of elementary particles (electrons, neutrons, etc.) and the forces between them, so cognition is nothing more than the activities of individual processing units in connectionist networks.

The problem is that the manner in which macrophysics is reducible to microphysics is not at all obvious. As traditionally construed, reducibility involves biconditional correlations (based on definition or law) between every reduced property and some reducing property. Unfortunately, despite the fact that many people are committed to the notion that microphysics is more fundamental than macrophysics, no one has been able to state any biconditional bridge laws which will actually effect the reduction of macrophysical properties to microphysical properties. In the absence of such laws it is very hard to see how the claim that the microphysical is more fundamental than the macrophysical can be justified. Indeed this state of affairs has led some philosophers to conclude that microphysics is not the more fundamental science (Horgan 1982). In any case, it is not at all obvious that likening the relationship of the symbolic to the subsymbolic to the relationship of the macrophysical to the microphysical will shed much light on how the symbolic is supposedly reducible to the subsymbolic.

Moreover, even supposing that the symbolic is reducible to the subsymbolic in the way that Smolensky suggests (that concepts literally *are* patterns over large numbers of subsymbols), it wouldn't automatically follow that the subconceptual is the correct level of explanation for cognitive phenomena. For, as Putnum (1980) has taught us, the correct level of explanation for a phenomenon is not always the same as the level of the basic entities which constitute the phenomenon. To use Putnum's well-worn example, an explanation of the fact that a cubical peg (one-sixteenth of an inch less than one inch wide) passes through a square hole (one inch wide) and not a round hole (one inch in diameter) is not to be found in the laws of particle mechanics and electrodynamics even if it is in some sense deducible from these laws. Rather, an explanation of the fact in question is to be found in certain laws of classical mechanics and geometry, viz., the board (with the holes in it) and the peg are rigid, the square hole is bigger than the peg, and the round hole is smaller than the peg. That is to say, higher level structures sometimes come under laws that are, in effect, autonomous from the laws describing their microstructure. This, of course, is exactly what Fodor (1975) and fellow travelers have in mind when they argue for the autonomy of the psychological (conceptual level). The upshot is that even supposing that the symbolic is reducible to the subsymbolic, the correct level for psychological explanation may not be at the level adopted by the subsymbolic paradigm (the subconceptual level). This in turn suggests that the subsymbolic account of cognition may be quite compatible with the symbolic account of cognition.

Smolensky denies this, however: He explicitly maintains the incompatibility of the symbolic and subsymbolic accounts of cognition. This suggests that he has in mind a notion of reduction which is much stronger than that reputed to hold between macrophysics and microphysics. For there must be a sense in which the higher conceptual level doesn't matter – a sense in which cognition can be completely explained away in terms of patterns of activity over large numbers of subsymbolic entities. In short, the claim that the subconceptual level is more fundamental than the conceptual level really amounts to the claim that there are no autonomous facts about cognition above the conceptual level.

This brings us to the relationship of the subconceptual level to the neural level. According to Smolensky, the relationship of the subsymbolic to the neurophysiological is such that "the best subsymbolic models of a cognitive process should one day be shown to be some reasonable higher-level approximation to the neural system supporting that process" (sect. 4, para. 10). Despite the fact that he does not mention it, this makes the relationship between the subsymbolic and the neurophysiological sound as much like the reputed relationship of the macrophysical to the microphysical as does the relationship of the symbolic to the subsymbolic. In this case, however, subsymbolic models must be taken as describing the macrostructure of (for lack of a better word) *sub*cognition (vs. the microstructure of cognition) and neural models must be taken as describing the microstructure of subcognition. Nevertheless, when it comes to the analysis of facts at the subconceptual level, the subconceptual level is supposed to be the more fundamental level. That is, although the ultimate constituents of subconceptual structures are neural, the correct level for explaining subconceptual facts does not, according to Smolensky, lie at the neural level. This does not seem to be an unreasonable position to hold – no more so than Putnum's claim that one cannot explain facts about square pegs and round holes in terms of microphysics, even though their ultimate constituents are microphysical.

Smolensky has in mind a claim that is much stronger than the claim that the subconceptual level is the correct level for the analysis of subconceptual facts. He also believes that the subconceptual level is the correct level for the analysis of conceptual (psychological) facts. As I have urged, such a claim cannot be justified by appealing to the reducibility of macrophysics to microphysics. What Smolensky needs is a relationship of reducibility between the symbolic and the subsymbolic in which the conceptual is not autonomous from the subconceptual. One possibility would of course be to treat the symbolic as some sort of logical construction out of the subsymbolic, much the same way numbers are sometimes treated as logical constructions out of sets. In the absence of a clearer understanding of "units" and "weights" (the basic subsymbolic entities), I cannot imagine what such an account would look like. However I fear it would come up against all the difficulties which have traditionally afflicted attempts (such as those of the logical behaviorists) to ground identities for mental entities in logic and set theory. In any case, until an account of the exact nature of the "reduction" of the symbolic to the subsymbolic is at least adumbrated, it is very difficult to evaluate the promise of PTC for philosophy of mind and cognitive science.

The psychological appeal of connectionism

Denise Dellarosa

Psychology Department, Yale University, New Haven, Conn. 06520

The appeal of connectionism has its roots in an idea that will not die. It is an idea that was championed by Berkeley, Hume, William James, Ebbinghaus, and (in a different form) the entire behaviorist school of psychology. Put simply, this idea is that cognition is characterized by the probabilistic construction and

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

activation of connections (or associations) among units: ideas (Hume), habits (James), words (Ebbinghaus), or stimulus–response pairs (behaviorism). It is also an idea that is represented in Smolensky's target article, albeit in distributed fashion.

The British empiricists and early American psychologists took great care to describe the essence of cognition as the building of associations through experience: Events that co-occur in space or time become connected in the mind. Events that share meaning or physical similarity become associated in the mind. Activation of one unit activates others to which it is linked, the degree of activation depending on the strength of association. This approach held great intuitive appeal for investigators of the mind because it seemed to capture the flavor of cognitive behaviors: When thinking, reasoning, or musing, one thought reminds us of others.

Historically, the crux of the issue has been whether these associations can be formalized best as a chain of pattern–action pairs linked together through inference (e.g., GPS: Newell & Simon 1972; Logic Theorist: Newell et al. 1958), or as a *network* whose units can be activated in parallel (e.g., Pandemonium: Selfridge 1959; HEARSAY: Reddy et al. 1973). Although symbolic rule-based models have had great success, there is a sense in which psychologists have never been quite satisfied with them as models of cognition, often turning them into hybrid models that include spreading activation networks (e.g., ACT*). [See Anderson: "Methodologies for Studying Human Knowledge" *BBS* 10(3) 1987.] Practically, as Smolensky points out, these models tend to suffer from an unwanted "brittleness": "best guesses" are difficult to achieve, and stimulus "noise" can bring operations to a grinding halt. Gone is the fluidity that flavors the human cognitive functioning observed in life and laboratory.

This is not to say, however, that symbolic models of cognition are worthless or simply false. Indeed, Smolensky is, I believe, right on target when he states that connectionist models stand to symbolic ones as quantum mechanics stands to classical mechanics. Just as the behavior of a physical system can be described using both classical and quantum terms, so too can the behavior of a cognitive system be described by both symbolic and connectionist models. In neither case, however, is the lower-level description a mere expansion of the higher, nor can a one-to-one mapping of constructs between the two be made. Moreover, just as quantum theory changed our thinking about the nature of physical systems and their fundamental processes, so too, I believe, are connectionist models challenging and changing our ideas about the nature of fundamental cognitive mechanisms. The most telling example is the treatment of inference by the two frameworks. It is a belief of many cognitive scientists (most notably, Fodor 1975) that the fundamental process of cognition is inference, a process to which symbolic modelling is particularly well suited. While Smolensky points out that statistical inference replaces logical inference in connectionist systems, he too continues to place inference at the heart of all cognitive activity. I believe that something more fundamental is taking place. In most connectionist models, the fundamental process of cognition is not inference, but is instead the (dear to the hearts of psychologists) activation of associated units in a network. Inference "emerges" as a system-level interpretation of this microlevel activity, but – when representations are distributed – no simple one-to-one mapping of activity patterns to symbols and inferences can be made. From this viewpoint, the fundamental process of cognition is the activation of associated units, and inference is a second-order process.

Certain connectionist models also challenge our understanding of cognition by representing symbols not as static data structures, but as activation patterns that occur momentarily at run time. Such dynamic, distributed instantiations of symbols hold great promise for the much-hoped-for marriage of cognitive science to neuroscience. For, although Smolensky takes great pains to explain that connectionist networks, as they presently stand, do *not* represent neural networks, it is exactly this type of distributed representation scheme that may be needed to explain how, for example, the same groups of neurons can be used to store a variety of memories in the brain.

Smolensky also discusses the ramifications of representing symbols in a distributed fashion in his reply to Pylyshyn (1984), suggesting that decontextualized symbols are a rarity – or impossibility – in connectionist models. This tends to give the knowledge encoded in connectionist networks a decidedly episodic flavor, a characteristic with great psychological significance. Much has been made in the psychological literature of the semantic knowledge/episodic knowledge distinction, where semantic knowledge is knowledge that is free of one's personal history. In reality, it is often difficult to uncover such decontextualized knowledge. Subjects' retrieval of facts are often spontaneously augmented by unbidden personal memories, such as when a fact was first learned, from whom it was learned, etc. Human knowledge seems to be cut from whole cloth, with fact and context inextricably interwoven.

The proof, however, still remains in the performance of these models as predictors of human behavior. As in the case of quantum versus classical mechanics, connectionist models must demonstrate a greater degree of precision and accuracy in predicting and explaining the many nuances of human behavior than symbolic models currently do. Their success, should it come, will mean a reinstatement of associationism as the cornerstone of cognition.

Some assumptions underlying Smolensky's treatment of connectionism

Eric Dietrich and Chris Fields

Knowledge Systems Group, Computing Research Laboratory, New Mexico State University, Las Cruces, N.Mex. 88003

Smolensky deftly avoids, in his target article, the issue of the scientific merit of, and hence the scientific evidence for, his particular formulation of the connectionist strategy in cognitive science (sect. 1.2). He rather attempts to define connectionism in a way that clearly sets it apart from the traditional symbolic methodology, and in so doing to argue obliquely for its superiority as a research strategy. Smolensky advances, in other words, a position in the philosophy of science, and it is in this spirit that we will reply.

The principle thesis of the target article is that connectionism – or at any rate Smolensky's formulation of it – is revolutionary in the sense that it is incompatible in principle with the received view, that is, the symbolic methodology. We will show that Smolensky's argument for this point, as presented in sections 2–5, rests on two implicit assumptions: the assumption that there is a "lowest" psychological level of analysis, and the assumption that different semantics, i.e., different interpretations of the behavior of a system, are in principle appropriate to different levels of analysis. It is of interest that Smolensky shares these assumptions with mentalists such as Fodor (1986) and Pylyshyn (1984), the very theorists he sees himself as opposing (sect. 1.3). We believe these assumptions to be false. Although we will not have space to argue in detail against them, we will illustrate briefly the effect of their rejection on the status and utility of the connectionist methodology.

Let us first examine the overall argument of the paper, taking Smolensky completely literally on each of his points. The goal of Smolensky's formulation of connectionism as stated in the Conclusion is not to replace symbolic cognitive science, but rather to enrich cognitive science as a whole by explaining "the strengths and weaknesses of existing symbolic theory . . . how symbolic computation can emerge out of nonsymbolic computation," and

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

so forth. Smolensky's goal, to use his analogy (sect. 5; also Smolensky 1987a), is to create a microtheory of cognition that stands to macroscopic cognitive science as quantum mechanics stands to classical mechanics. Based on this analogy, one might expect Smolensky to propose a microtheory, the explanations of which grade smoothly into those of the macrotheory as a set of parameters approach specified limits. Instead, Smolensky argues at length that the micro- and macrotheories are in this case inconsistent (sect. 2.4), and indeed, he phrases the argument in a way that suggests that he believes the scientific utility of connectionism to hinge on its being inconsistent with the symbolic approach (hence his impatience with "bland ecumenicalism").

The notion that one theory can explain the strengths and weaknesses of another theory with which it is flatly inconsistent is perplexing, to say the least. The way around the paradox, clearly enough, is to view the theories as advancing alternative interpretations of the system's behavior to satisfy different explanatory requirements – in the case of physics, classical mechanics explains billiard balls, while quantum mechanics explains electrons. The interfaces are then handled by defining approximations, This is what Smolensky does in practice (sect. 5), but it is straightforwardly ecumenical. Smolensky's attempt to advance this (quite reasonable) view of the relationship between the symbolic and subsymbolic approaches while simultaneously affirming their inconsistency in principle lends his paper a certain dramatic tension, but hardly increases its coherence.

Smolensky's insistence that the symbolic and subsymbolic approaches are inconsistent can be traced, we believe, to his assumptions about the role of semantics in psychological explanation. Before proceeding with this, however, it is worth making fully explicit a point that Smolensky touches on, but does not elaborate. The philosophical debate about connectionism is not, contrary to common opinion, a debate about architecture; it is only a debate about semantics, that is, about the interpretation of the behavior of an architecture. This point can be seen clearly by viewing the connectionist architecture in the state-space representation defined by Smolensky's claim (8a). The states in this space are vectors specifying possible values of the excitations of the nodes in the system; this space is continuous, but can be approximated arbitrarily well by a discrete space (sect. 8.1). Paths in this state space correspond to episodes of execution from given inputs, and can be viewed as searches in the usual way.

Viewed in this representation, computations on a connectionist machine have the same architectural features as computations on a von Neumann machine; they amount to serial searches in a space of possible solutions. In particular, a deterministic connection machine is just as behaviorally rigid, and hence just as brittle, as a deterministic von Neumann machine. This equivalence is preserved if the machines in question are stochastic (Fields & Dietrich 1987). As Smolensky notes, somewhat obliquely, in sect. 2.4, what connectionism has to offer architecturally is no more, but also no less, than an alternative methodology for building AI systems.

Smolensky's revolution can therefore only be supported by demonstrating a principled incompatibility between the interpretations of the behavior of the architecture advanced by his version of connectionism on the one hand, and by the symbolic paradigm on the other. The entire weight of Smolensky's case, therefore, must rest on claims (8b) and (8c), which together amount to the claim that such a principled incompatibility of interpretations exists. Claim (8b) is simply asserted; Smolensky then argues that, if (8b) is true, it will typically be impractical to calculate the interactions between patterns of activity interpretable at the conceptual level precisely. The consequence of this argument is then reformulated from a claim about computational resources to a principle in (8c). This move deserves some skepticism: Given a mapping from activation patterns to concepts, and a precise specification of the underlying dynamics, one could calculate a precise specification – or at least an arbitrarily good approximation (e.g. to within 10^{-16} seconds) – of the behavior of the system at the level of concepts if one were willing to take the time to do so. Smolensky grants us a precise specification of the underlying dynamics; indeed this is, one suspects, what makes connectionism dear to him. He must therefore implicitly deny that any fixed mapping from activity patterns to concepts is possible, in principle.

Two questions arise very naturally at this point. First, how can Smolensky rule out the simple stipulation, in the spirit of denotational semantics, of a fixed interpretation mapping activity patterns to concepts? Humans, after all, interpret human behavior conceptually with great facility, and there is every reason to believe that they do not do so by calculating approximate concept-level descriptions from the underlying dynamics. What is to prevent theorists from doing the same? Second, if there is something to the claim that concept-level descriptions are fuzzy in principle, what prevents us from using the same argument to show that subsymbolic descriptions are only fuzzy approximations of neural descriptions, or that neural descriptions are only fuzzy approximations of biochemical descriptions, and so forth? The answers to these questions reveal Smolensky's implicit assumptions, which he appears to share with his arch-rivals Fodor and Pylyshyn.

Let us consider the second question first. Smolensky argues (sect. 4) that connectionists need not be too concerned with neural realism because we do not currently know enough about the brain to construct neurally realistic models. We must admit finding this claim somewhat mystifying in light of the success of neurally minded groups such as Grossberg's in formulating candidate models of interesting cognitive processes (e.g., Grossberg 1980, 1987; Grossberg & Mingolla 1985; Grossberg & Stone 1986). Be that as it may, however, it is surely an error to argue from impatience over a lack of data to a principled distinction between levels of description. Smolensky must, in addition, implicitly believe that there is no psychological description of events at the neural level for subsymbolic descriptions to be fuzzy approximations of. In other words, if the subsymbolic level of description is the lowest level that admits a psychological interpretation, then we can stipulate that descriptions at this level are precise, and we need not worry that they may turn out to be mere fuzzy approximations of lower-level descriptions.

A consideration of the first question posed above reveals a second assumption. Smolensky cannot block the claim that the behavior of a physical system can be interpreted, by stipulation, at any level of description that its interpreters prefer. The model-theoretic notion of interpretation is, after all, the very cornerstone of the theory of virtual machines on which the practice of emulative programming rests. He must assume, therefore, that some interpretations are in principle appropriate to the chosen level of description, whereas others are not. Claim (8b) is an example of such an assumption; in addition to (8b), Smolensky must assume that the rough and ready conceptual interpretations that humans impose both on themselves and on each other in everyday life are inappropriate as high-level descriptions of connectionist systems that are precise relative to the operative explanatory goals. Apparently for Smolensky, the appropriateness of a description is determined not by explanatory goals, but by metaphysics.

"Thoroughly modern mentalists" such as Fodor and Pylyshyn make precisely these assumptions, although with the conceptual level taken to be the "preferred" level of analysis (e.g., Pylyshyn 1984; Fodor 1986). Smolensky simply transfers this preferential treatment down one level of description, while maintaining the same mentalist, or we shudder to say, dualist metaphysics.

If these assumptions are rejected, one is left with an ecumenical, but in our opinion far from bland theory. It runs like

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

this: Humans are complex information-processing systems. They can be interpreted as such at any level of description. A description of a human as a dynamical system at any level can be used to calculate a description at any higher level; if one is willing to commit the necessary resources, this description can be made arbitrarily precise relative to the lower-level description. The precision required is determined by explanatory goals. The only restrictions on the semantics of the interpretations used to describe the system are those imposed by intra-level coherence, and by the explanatory goals with which the interpretation is constructed. One can, if one wants, interpret neurons as representing grandmothers; if this interpretation does not prove to be useful, it can always be revised.

We wish to claim no credit whatsoever for this view. It was formulated over 30 years ago by Ross Ashby (1952), and it appears to us to provide, with relatively minor technical embellishments, a quite adequate foundation for computational psychology and cognitive science. In particular, it shows us clearly how connectionism, viewed not as a revolution, but as a valuable addition to our methodological tools, can achieve the goals Smolensky sets out in his Conclusion.

On the proper treatment of Smolensky

Hubert L. Dreyfus[a] and Stuart E. Dreyfus[b]

[a]Department of Philosophy, University of California, Berkeley, Calif. 94720 and [b]Department of Industrial Engineering and Operations Research, University of California, Berkeley, Calif. 94720

Connectionism can be understood as the most serious challenge to representationalism to have emerged on the cognitive science scene. In view of the intellectual rigor of Smolensky's contribution to connectionism we hoped his target article would present a powerful formulation of this eliminitivist challenge. On first reading, however, we were shocked by what seemed to be an attempt to defend a two-level *representationalist* account of connectionism. At one level – call it the macrolevel – Smolensky seemed to be saying that whenever a net is processing information the representational symbols of conventional cognitive science are instantiated by patterns of activities of units. His disagreement with cognitive science is simply that the generation of intelligent behavior could not be exhaustively explained by formal operations on these "conceptual" symbols. Moreover, Smolensky seemed to hold that on a second, deeper microlevel, one could carry out the conventional cognitive science program by explaining intelligent behavior using finer-grained (subconceptual) symbols which picked out context-free microfeatures of the task domain.

That this was not only our interpretation was brought home to us when we found that Fodor & Pylyshyn (1988) quote this very target article as clearly placing Smolensky and the connectionists in the representationalist camp. Fodor & Pylyshyn use Smolensky's statement that "Entities that are typically represented in the symbolic paradigm by .symbols are typically represented in the subsymbolic paradigm by a large number of subsymbols" (sect. 1.3., para. 5) as evidence that Smolensky is committed to what we have called macrorepresentationalism. To support their understanding that Smolensky is a representationalist at the microlevel Fodor & Pylyshyn quote Smolensky's connectionist hypothesis that, "Complete, formal and precise descriptions of the intuitive processor are generally tractable not at the conceptual level, but only at the subconceptual level" (8c). Fodor & Pylyshyn conclude that "the resultant account would be very close to the sort of language of thought theory suggested in early proposals by Katz and Fodor."

On second reading we think that we (and Fodor & Pylyshyn) have misunderstood Smolensky on both these points. There is a plausible way of construing both statements, and others like them, in the context of the whole target article which makes clear that Smolensky's version of connectionism is not committed to representationalism on either the macro- or the microlevel.

On the macrolevel, the sentence quoted by Fodor & Pylyshyn does seem to endorse the cognitivist hypothesis that all intelligent behavior can be analyzed as the sequential transformation of symbols which represent context-free features of the object domain – precisely those features we can normally notice and name. However, in the course of Smolensky's paper the same idea is progressively refined until it is clear that the claim is not that *all* intelligent behavior involves symbol transformation, but rather that only a very limited form of behavior – the deliberate behavior typical of the novice consciously applying rules – involves symbols. Hypothesis 8b restates the same principle but explicitly refers to "the semantics of *conscious* concepts of the task domain" (our emphasis). And in section 6.1, paragraph 1, Smolensky reinterprets 8b in a completely unambiguous way stating that "concepts that are *consciously accessible* correspond to patterns over large numbers of units" (our emphasis).

On the microlevel, Smolensky certainly seems to be a representationalist when he says in his abstract: "The numerical variables in the system correspond semantically to fine-grained features below the level of the concepts consciously used to describe the task domain." And the final statement of hypothesis 8 might well suggest, and indeed did suggest to Fodor & Pylyshyn, that the precise logical formalism they favor, although it is missing on the macrolevel, which presupposes a language of thought using features of the sort named in everyday language, can be found on the microlevel in a language of thought that uses subsymbols, standing for microfeatures that we are not normally able to perceive and articulate.

Once one realizes, however, that what Smolensky means by a complete, formal, and precise description is not the logical manipulation of context-free primitives – symbols that refer to features of the domain regardless of the context in which those features appear – but rather the mathematical description of an evolving dynamic system, it is far from obvious that the fine-grained features Smolensky calls subsymbols are the elements of a language of thought. Sentences such as "Hidden units support internal representations of elements of the problem domain, and networks that train their hidden units are in effect learning subconceptual representations of the domain" (sect. 3, para. 5) certainly leave open the possibility that there necessarily exist context-free subsymbols representing features of any problem domain, and that units or patterns of units detect them. But later in the paper Smolensky forecloses this interpretation by stating explicitly that "the activities of the subconceptual units that comprise the symbol – its *subsymbols* – change across contexts" (sect. 7.2., para. 6).

If the idea of a unit having a semantics but not corresponding to a context-free feature of the task domain seems contradictory, consider the following. Given a net with hidden units and given a particular activity level of a particular hidden unit, one can identify every input vector which produces that activity level of that hidden unit. The activity level of the unit can then be semantically interpreted as representing the set of these input vectors. That unit at that activity level is a subsymbol in a very weak sense. For, while such a subsymbol can be correctly interpreted as representing a microfeature of the domain, the microfeature need not be a context-free microfeature. (To symbolize a context-free microfeature the unit would need to take on its given activity level independently of one or more of the elements of the input vector. Roughly put, if the input vector is the state of the world, then a context-free feature is that portion of the input vector which determines the activity of the hidden unit independent of the rest of the input vector). Given this weaker version of semantic interpretation, there is no necessary connection between claiming that hidden units are semantically

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

interpretable, as Smolensky holds, and claiming that they pick out context-free invariant features of the domain, which is the implicit commitment of the representationalism characteristic of cognitive science. Thus connectionists can hold the minimal representationalist position of Smolensky and still eliminate the sort of conceptual and subconceptual symbols defended by cognitivists such as Fodor & Pylyshyn.

The promise and problems of connectionism

Michael G. Dyer

Computer Science Department, University of California at Los Angeles, Los Angeles, Calif. 90024

I am attracted by PDP (parallel distributed processing) models because of their potential in tolerating noise, generalizing over novel inputs, exhibiting rule-like behavior without explicit rules, performing pattern completion and decision-making through massively parallel constraint satisfaction, and so on. Connectionism promises to supply an alternative foundation for cognitive science (in place of the symbolic, linguistic, logical foundations) and to unite it, for the first time, with the physical and biological sciences through statistical and continuous models.

The use of weight matrices does allow widely varying inputs to be mapped into a set of invariant categories in the output. Unfortunately, what gives these systems their robustness also makes it very difficult to capture abstractions and symbolic operations and has led to attacks on the claims made for specific PDP models (e.g., Pinker & Prince 1988).

Although the discovery of automatic learning devices may appear to supply a philospher's stone to cognitive science, such devices must still be programmed (by specifying the initial connections of the network and the nature and order of the input during the learning phase). Although it might be theoretically possible to submit enormous quantities of carefully organized input data to one gigantic, homogeneous "connectoplasm" and after 20 years to get out a college-educated neural network, this would be impractical, to say the least. Even if it were successful we wouldn't understand scientifically what we had produced. Consequently, we must also pursue top-down approaches. Higher-level tasks must be specified and used to direct the construction of PDP architectures capable of handling those tasks, which invariably require symbol-like operations. Consider reference resolution during comprehension and rebuttal during argumentation:

(1) Pronoun reference. To read the text <John walked into a restaurant and the waiter, Bill, walked up. After he ordered he brought him the food.> symbolic NLP (Natural Language Processing) systems first instantiate a restaurant schema and bind John to the patron role. Within the schema is a representation for the patron receiving the food from the waiter. Since John is already bound to the patron role, upon binding "him" as the recipient of the food, the system can infer immediately that John (rather than Bill) is receiving the food. This fundamental kind of role-binding (Dolan & Dyer 1987) is very difficult to accomplish in a reasonable way in PDP models.

(2) Rebuttal. If one goes up to a Finnish friend and states: "All Finns are terrible at music," the chances are good that he will reply: "What about Sibelius?" Consider for a moment what must be going on here. [See also *BBS* multiple book review of Sperber & Wilson's *Relevance*, *BBS* 10(4) 1987.] First, the Finn must understand the initial utterance and realize that it is not a fact but an opinion or belief. He must then decide whether or not he agrees with it. If he has a negation of it already in memory, then he need only recall it. However, it is not likely that someone has even expressed this thought to him in the past, so a negation has to be generated on the fly. He needs to use an argument strategy (Alvardo et al. 1986) of the sort: <if x claims that all y from class C have property P, then search in memory for a y′ from C with not-P>. Finally, the proper associative retrieval has to be performed and the recalled counterexamples have to be evaluated before actually being generated as a rebuttal. The above task is complex, involves a number of steps, and cannot be modeled with only simple associational techniques (which, solely based on the input, would first tend to retrieve all Finns who are terrible musicians).

When association lists were first implemented in Lisp (as P-lists), researchers argued that all forms of knowledge could be represented in terms of such associations. But P-lists (frames, etc.) are only tools. They do not tell us how to construct any particular theory of cognitive processing. The invention of Lisp greatly advanced the technological base for pursuing symbolic AI. Now researchers are realizing that all forms of associations can be represented as adaptive weight matrices. The discovery of the generality and usefulness of adaptive weight matrices may advance the technological base for modeling one kind of learning just as greatly. But, like other tools, adaptive weight matrices do not supply us directly with solutions to complex processing problems in cognitive modeling.

Symbolic operations *should not be* implemented to perform exactly as in symbolic models. If they were, PDP models would lose many of their interesting properties. But without mechanisms to perform *analogous* symbolic functions, PDP models will never advance beyond the signal processing stage. Without something analogous to bindings, it is impossible to form larger knowledge structures.

Currently, schema-like knowledge is modeled in PDP systems in terms of patterns of activation over PDP units, usually composed of "microfeatures" (Rumelhart & McClelland 1986). This approach poses a number of problems for anyone who wants to represent and manipulate schema-like structures for comprehension, question answering, argumentation, and so on. These problems include:

(1) Knowledge portability: In a symbolic system, any procedure that knows the syntax of the symbolic formalism can execute or interpret a symbol's semantic content. Symbol structures thus serve as an interlingua for internal processes. In contrast, since PDP models form their own patterns of activity through learning, the activity pattern learned by one network will generally be undecipherable to another. As a result, it is very difficult to port knowledge from one area of memory to another. Most current PDP models are designed to perform a single task; the same network cannot be used for multiple tasks.

(2) Microfeature selection: In many PDP models, gradient descent learning leaves a set of microfeatures "clamped down" or fixed. In such cases, the knowledge representation problem of AI is simply pushed back to an equally hard problem of determining what the microfeatures must be. One way to eliminate microfeatures is to extend gradient descent learning techniques into the representation of the input itself. In one experiment, our model (Miikkulainen & Dyer 1987) learned distributed representations of words while at the same time performing McClelland and Kawamoto's (1986) case-role assignment task. As a result, microfeatures were never needed; word meanings were formed distributively during the performance of the task, and the resulting model did a better job at generalizing to novel inputs in the case-role assignment task.

(3) Training set reuse: In most PDP models there is no distinction between rapidly changing knowledge and knowledge more impervious to change. The weight matrices are formed by repeated interaction with a training set (TS). Since the same network stores all associations, the teacher must resubmit the original TS to the system when novel inputs are learned; otherwise the system will not respond correctly to the old inputs (due to interference). In contrast, humans can often form categories based on a very small number of inputs (Pazzani & Dyer 1987). It is unrealistic to assume that TS data are stored

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

verbatim somewhere in the brain, simply to reestablish a weight matrix. Clearly, knowledge must be consolidated at some point and memories must be organized so that new information can be rapidly acquired while interfering minimally with consolidated memories.

PDP models have caused great excitement, especially in signal processing areas (where adaptive networks allow complex data to be mapped to a fixed set of categories). If PDP models are to go beyond adaptive signal processing, however, analogs for symbols and symbol processing (Touretzky 1987) will have to be found.

Dynamic systems and the "subsymbolic level"

Walter J. Freeman

Department of Physiology-Anatomy, University of California, Berkeley, Calif. 94720

I find it easy to agree with the premises that Smolensky sets forth in defining and defending the subconceptual hypothesis, and with many of the conclusions that he draws. In particular, the observations by my students and myself on the functions of the brains of small mammals trained to perform cognitive tasks appropriate to their stations in the phylogenetic tree have amply shown that (8a) the dynamics of intuitive processors and their changes with learning are governed by ordinary differential equations (ODEs), that (8b) concepts are complex patterns of activity over many units, and that (8c) formal, precise and complete descriptions are not "tractable" at the conceptual level, but only at the subconceptual level.

But my agreement is predicated on his agreement with me that formal, complete, and precise descriptions refer to the ODEs by which a model is formulated, whether in software or hardware, and not to the solutions of the equations, which with any reasonable degree of model complexity are endlessly unfolding, evolving, and delightfully (or painfully) full of surprises. Given the structures of the ODEs that suffice to replicate EEG waves and the requisite complete parameter values, we are no more able to predict in detail the output of the model than that of the modeled brain (Skarda & Freeman 1987). Will he agree with me that the solutions may constitute concepts, exemplified by odors (meaningful events for an animal) as distinct from odorants (laboratory chemicals used as conditioned stimuli)? This is providing that the solutions conform to stable attractors representing convergence to reproducible spatiotemporal patterns of neural or simulated activity (Baird 1986; Freeman & Skarda 1985).

With a similar reservation I agree that the subconceptual is above the neural level. By this I mean that information which relates to or subserves goal-directed animal behavior exists at the macroscopic level of cooperative activity of masses of neurons; it is not observable in the unaveraged behavior of single neurons (Freeman 1975; Freeman & Skarda 1985; Skarda & Freeman 1987). Most of the physiological results that claim otherwise have been derived from paralyzed or anesthetized animals, or they have resulted from stimulus-locked time ensemble averaging, which retrieves the 2% of variance related to the stimulus and flushes the 98% of the background noise that arises from the attractors in the brain. I also agree that the proper time base for solving descriptive ODEs is continuous, but with the understanding that in software simulation a discrete time step is needed, small enough to be nonintrusive; this is feasible. But in hardware simulation too the array must be small enough (e.g., up to 100) so that one need not resort to time multiplexing; with any reasonable size of fully connected array (e.g., 1,000 to 10,000 elements), continuous time is not feasible. This is because without multiplexing the number of connections increases with the square of the number of elements, N, but with multiplexing it increases as 2N (Freeman, in press).

I presume that Smolensky shares my aversion to partial differential equations, perhaps for the same reasons, that they are infinitely dimensional and less tractable than integro-differential equations. The nervous system also appears to avoid them at the level of concept formation by recourse to spatial discretization with columns and glomeruli. I suppose that by "quasilinear" equation Smolensky means the 1st or 2nd order ODE cascaded into a sigmoid nonlinearity or its functional approximation. This element for integration is by now virtually standard in connectionism.

I agree most enthusiastically with his complaint that altogether too great a proportion of our understanding of real brains is structural rather than functional. By the same token I complain that connectionists likewise ignore dynamics too readily; those who do pay attention tend to rely altogether too greatly on equilibrium attractors for their dynamics and neglect the attractors of limit cycles and chaos. The reasons for both instances of neglect appear to be the same: Things are too complicated.

Smolensky's Table 1 should, in my opinion, have all pluses; this can be done with some minor redefinition of terms.

I infer, from his reliance on nonlinear ODEs, that Smolensky agrees that solution sets for a complex network typically incorporate multiple stable domains. The behavior of such a system is most often marked by sudden jumps or bifurcations from one domain to another, depending on one or another bifurcation parameter, which are the phase transitions (sect. 9.5., para. 1.) that he writes about. Although the local time scale may be continuous, the global time scale must be discrete. This is the essence of the strings of bursts that we see in EEGs and the strings of concepts that we infer the bursts carry (Skarda & Freeman 1987), which are parts of the "slowly shifting harmony landscapes" in his language.

Smolensky asks questions of doubtful value concerning the semantics of subsymbolic systems, namely, "which activity patterns actually correspond to particular concepts, or elements of the problem domain?" Animals, of course, have no proper linguistic abilities, and we have as yet no appropriate data from man. Correspondence is a matter of behavioral correlation in neurophysiological research. We have made extensive use of the third class of methodologies that Smolensky exemplifies by "multidimensional scaling." The first two are laden with difficulties that we believe are insoluble. The implicit notion that any positivistic relationship exists between "words," "images," and "neural activity patterns" should have died with phrenology but regrettably has not.

Moreover, the idea of representation carries with it notions of registration, storage, retrieval, comparison, cross-correlation, and figure completion. The PDP (parallel distributed processing) operations of backward propagation, error correction, and heteroassociative learning are also predicated on this digital-computer-metaphor for memory. Our physiological evidence shows that such storage, retrieval, comparison, etc., do not exist in the olfactory system, and we have predicted that they will be found not to exist elsewhere in biological brains. We know that information is incorporated into the dynamics and the structure of brains from the outside world, but we do not know, nor need we know, what is being represented on which TV screen to which homunculus in the brains of rabbits, or, for that matter, of our spouses. The idea is unnecessary in understanding brains and the devices that simulate them. Fodor and Pylyshyn (1987) are, I believe, right in stating that if connectionism relies on representation it is dead. Its liveliness stems from its independence of that cognitivist assumption.

I much prefer Smolensky's "harmony maxima" to his "veridical representations." The latter term has the ring about it of "truth tables," which exist in the world as teeth and fire but not in the brain; what evidence can he put forth that any one of them

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

is true? But the former has the ring of poetry, which Turing machines cannot do or even translate, and which lends itself to the images of strange attractors and the genesis of chaos, the grist and gristle of dynamical systems.

Connectionism and the study of language

R. Freidin

Department of Philosophy, Princeton University, Princeton, N.J. 08544

The application of computers to the problem of understanding natural language has, from the outset, been marked with great optimism and also great naivete (Dresher & Hornstein 1976). This seems to be no less true for current connectionist approaches to language – though it is perhaps a bit early in the game to see what connectionist models can do in the area of natural language. (See, however, Pinker & Prince 1987 for a detailed critique of Rummelhart & McClelland's 1986 proposal regarding the past tenses of English verbs.) What characterizes these AI approaches to natural language is a certain lack of understanding about the complexity of the object of inquiry and the difficulty of the problem. For example, Smolensky states in sect. 6, para. 4 that "the competence to represent and process linguistic structures in a native language is a competence of the human intuitive processor, so the subsymbolic paradigm assumes that this competence can be modeled in a subconceptual connectionist dynamical system." By competence Smolensky means ability, I assume. He is therefore proposing a connectionist model of *linguistic performance* (not to be confused with *linguistic competence* – Chomsky's technical term for a speaker's knowledge of language). There is no reason to believe that such a model will succeed.

The most difficult problem a model of language use must address is what is called *the creative aspect of language use* – the fact that normal language use is innovative, potentially unbounded in scope, and free from the control of detectable stimuli. As Chomsky notes in *Language and Mind* (1972), the latter two properties could be accommodated within mechanical explanation. He continues:

> And Cartesian discussion of the limits of mechanical explanation therefore took note of a third property of the normal use of language, namely its coherence and its "appropriateness to the situation" – which of course is an entirely different matter from control by external stimuli. Just what "appropriateness" and "coherence" may consist in we cannot say in any clear or definite way, but there is no doubt that these are meaningful concepts. We can distinguish normal use of language from the ravings of a maniac or the output of a computer with a random element.
>
> Honesty forces us to admit that we are as far today as Descartes was three centuries ago from understanding just what enables a human to speak in a way that is innovative, free from stimulus control, and also appropriate and coherent. This is a serious problem that the psychologist and biologist must ultimately face and that cannot be talked out of existence by invoking "habit" or "conditioning" or "natural selection." (pp. 12–13)

Or "subconceptual connectionist dynamical system," it would appear.

The underlying assumption of connectionist approaches to cognitive modeling seems to be that we now have a line on the right architecture for cognition (i.e., the "connection machine" hardware) as well as the appropriate mechanism for learning (the software for running the machine successfully in cognitive domains). Presumably if we feed the connection machine the appropriate data, the machine will produce the correct cognitive model. The machine then functions as a discovery procedure for cognitive models in various domains. Thus in the domain of language we might expect that on presentation of data the machine will produce a grammar of the language.

The failure of discovery procedures for grammars is well known in linguistics. It seems highly unlikely that this situation will change with the introduction of some powerful computer architecture coupled with general inductive learning strategies of the sort discussed in the connectionist literature. What linguistic research into the structure of language has shown over the past thirty years is that a cognitive model of language involves several abstract concepts specific to the language faculty. This model has strong empirical support crosslinguistically (for details, see Chomsky 1984; Freidin 1987). There is no reason to believe that these abstract concepts will emerge from the kind of statistical analysis of data available in connectionist networks – either symbolic or subsymbolic.

In fact, there is good reason to believe just the opposite. Many of the abstract concepts of linguistic theory are embedded in the formulation of general grammatical principles which distinguish ungrammatical sentences from grammatical sentences. Consider, for example, the well-formedness conditions on the occurrence of bound anaphors (e.g., reflexive pronouns and reciprocals – *each other* in English). In current versions of generative grammar, there is a general condition called Principle A of the Binding Theory that prohibits anaphors that are antecedent-free in the domain of a syntactic subject (details aside). The effect of this principle is to mark sentences such as (1) as ungrammatical in contrast to grammatical sentences as in (2).

(1) *Mary$_i$ believes [$_S$ Bill to like herself$_i$]

(2) a. Mary$_i$ believes [$_S$herself$_i$ to like Bill]
    b. Mary$_i$ likes herself$_i$

The salient point here is that principles of grammar like Principle A are formulated on the basis of what structures are ill-formed – that is, in terms of ungrammatical examples which are not part of the normal linguistic environment. Thus the concepts involved in such principles, not to mention the actual formulation of the principles themselves, are motivated in terms of the "poverty of the stimulus" – that is, the lack of relevant information in the environment of the language learner (see Chomsky 1980 for discussion). The problem for any model of language acquisition based solely on input from the linguistic environment is that there is no way to distinguish ungrammatical sentences, e.g. (1), from novel grammatical sentences. Why should a language learner who has heard the sentences in (2) judge (1) to be ungrammatical rather than just a novel grammatical sentence? The answer (according to generative grammar) is that principles like Principle A are part of the innate cognitive structure a child brings to the task of language acquisition.

For such principles (or their effects) to be derived from connectionist networks constructed solely from the statistical analysis of data, the ungrammatical versus novel grammatical sentence problem must be solved. It is difficult to see how this is to be done without incorporating some version of the innateness hypothesis in linguistics. Furthermore, caution seems advisable when interpreting the effects of connectionist networks. For example, Hanson and Kegl (1987) discuss a connectionist network PARSNIP "that learns natural language grammar from exposure to natural language sentences" (from the title of the paper). On the evidence they present, the claim is false. "PARSNIP correctly reproduces test sentences reflecting one level deep center-embedded patterns (e.g., [[*the boy* [*the dog bit*]] *yelled*] which it has never seen before while failing to reproduce multiply center-embedded patterns (e.g., [[*the boy* [*the dog* [*the cat scratched*]*bit*]] *yelled*])." However, the phrase structure rules for English do not make distinctions between multiple center-embedding and single center-embedding – or for that matter, between center-embedding and noncenter-embedding. The unacceptability of multiply center-embedded constructions does not follow from grammar at all – see Miller and Chomsky (1963) for the original discussion. Thus, whatever the PARSNIP network represents, it is obviously not a grammar in

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

the usual sense. At this point is it still far from clear what relevance connectionist models are going to have to cognitive models of language.

Statistical rationality

Richard M. Golden

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

Because Smolensky's subsymbolic hypothesis requires a more rigorous formulation, his arguments are not convincing. In his commentary, a revised form of Smolensky's subsymbolic hypothesis is proposed based upon analyses of the relationships between logical inference, statistical inference, and connectionist systems. The compatibility of the symbolic and subsymbolic paradigms is then reconsidered using the revised subsymbolic hypothesis.

Problems with the terminology of Smolensky's subsymbolic hypothesis. Smolensky's subconceptual-level hypothesis (8c) is completely dependent upon distinguishing the "conceptual" and "subconceptual" levels of description, yet he is unable to characterize even the nature of the representation at the subconceptual level. The term "complete, formal, and precise description" in (8c) is also problematic. I believe that Smolensky's intention here is to describe the computational goals of a connectionist model with respect to his statistical "best fit principle." (sect. 9.1.) If this is the case, however, this intention should be explicitly stated within the subsymbolic hypothesis.

Logical inference is a special case of statistical inference. One serious limitation of Boolean algebra or symbolic logic is that propositions are either true or false. That is, symbolic logic is incapable of precisely representing partial belief in a proposition. Also note that the traditional rule-governed approach in cognitive science is based upon deciding whether propositions are either true or false: A proposition cannot be both "almost" true and "almost not" false. A number of statisticians (e.g., Cox 1946; Jeffreys 1983; Savage 1971) have proved that the *only* and *most general* calculus of partial belief whose conclusions are always guaranteed to be consistent with symbolic logic is probability theory.

Cox's (1946) approach is particularly elegant. Let α and β be propositions and let B $(\alpha|\beta)$ be a belief function whose value is one (true) if the truth of β implies the truth of α, and whose value is zero (false) otherwise. Thus, when the range of the belief function, B, is binary and discrete, B represents a *rule*. The problem now is to extend the range of the function B so that it is *continuous* and is permitted to range between zero and one inclusively. Cox (1946) has provided a simple formal argument showing that if B always assigns real numbers to pairs of propositions such that the laws of symbolic logic are never violated, the function $B(\alpha|\beta)$ must be the conditional probability of α given that β is true.

Rational connectionist models are statistical inference mechanisms. Although Smolensky has proposed the "best fit principle" as a desired property of connectionist systems, this principle has only been formally demonstrated for a small class of connectionist systems such as the Harmony theory neural networks of Smolensky (1986a) or the Boltzmann machine neural networks of Hinton and Sejnowski (1986). Other researchers (e.g., Hummel & Zucker 1983; Rumelhart, Smolensky, McClelland, & Hinton 1986) have viewed their networks as *constraint satisfaction* networks or nonlinear (usually quadratic) optimization algorithms (see e.g., Luenberger 1984) that are minimizing/maximizing some cost function. From the perspective of demonstrating rational information processing, such constraint satisfaction analyses are inadequate since the inference process has not been shown to be either logical or statistical in nature.

One solution to the problem of demonstrating *rational* information processing is to use the cost function that a neural network is minimizing/maximizing during information retrieval to construct a probability function that the network is maximizing during information retrieval (Golden, submitted). Moreover, it can be shown using an extension of arguments by Smolensky (1986a) that such a construction is unique (Golden, submitted). If such a probability function exists, then the neural network can be viewed as a statistical pattern recognition algorithm that is computing the *most probable* value of the information to be retrieved. Such algorithms are known as MAP (maximum a posteriori) estimation algorithms in the engineering literature. This type of computational theory provides formal justification for the "statistical rationality" of many popular deterministic connectionist models such as Anderson's BSB model (Anderson et al. 1977), Hopfield's (1984) model, and the back-propagation neural network models (Rumelhart, Hinton & Williams 1986). In addition, questions concerning what classes of probabilistic environments a given connectionist model is capable of learning and the extent to which a given connectionist model's learning algorithm is optimal can be addressed.

Continuity is necessary for representing partial (real-valued) beliefs. As noted above, statistical inference differs from logical inference in its ability to represent and manipulate "partial beliefs" in propositions. Logical inference can only *approximately* model statistical inference, but statistical inference can yield *exactly* the same answers as logical inference. Accordingly, the following revised version of Smolensky's subsymbolic hypothesis is suggested:

> **A revised subsymbolic hypothesis.** The intuitive processor is a connectionist dynamical system that is designed to solve statistical pattern recognition problems.

This revised subsymbolic hypothesis demonstrates more directly the incompatibility of the subsymbolic and symbolic paradigms as described by Smolensky in the target article. The reason why Smolensky's hypothesis (10) must be rejected is that, according to this revised hypothesis, the intuitive processor is representing and manipulating "partial beliefs" (i.e., belief functions whose range is *continuous* and not discrete) which cannot be done by a rule-governed processor. Note that the role of continuity in the connectionist paradigm was also stressed in Smolensky's target article (sect. 8.1.).

Conclusion. In summary, Smolensky's original subsymbolic hypothesis is too dependent on a characterization of the elusive "subconceptual" level of symbolic processing, and should instead stress the role of statistical inference in connectionist systems. The "continuous" aspect of statistical information processing relative to discrete logical information processing can then be used to prove the incompatibility of the symbolic and subsymbolic paradigms.

ACKNOWLEDGMENTS

This research was supported in part by the Mellon Foundation while the author was an Andrew Mellon Fellow in the Psychology Department at the University of Pittsburgh, and partly by the Office of Naval Research under Contract No. N-0014-86-K-0107 to Walter Schneider.

Common sense and conceptual halos

Douglas R. Hofstadter

Psychology and Cognitive Science, University of Michigan, Ann Arbor, Mich. 48104

Paul Smolensky's target article is an excellent clarification of the position of the connectionist movement in cognitive science. Since I agree with all its major points, I would like to take the opportunity to cast these issues in a somewhat different light. I believe that understanding and explaining the elusive nature of

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

common sense will become central concerns of cognitive science, if they are not so already. I will therefore attempt to draw some links between connectionism and common sense.

In his one allusion to common sense (sect. 2.3, hypothesis 5b), Smolensky writes: "The process of articulating expert knowledge in rules seems impractical for many important domains (e.g., common sense)"; this seems to suggest that common sense is a domain. I strongly believe such a suggestion should be marked "to be rejected."

"Common sense" is a term used frequently in discussions of what's wrong with artificial intelligence, yet is seldom if ever defined. I have heard it suggested that if an AI system absorbed vast amounts of knowledge (e.g., the *Encyclopaedia Britannica*), it would possess common sense. I believe this is totally wrong. An idiot savant who could recite the entire *Encyclopaedia Britannica* by memory would almost surely possess less common sense than ordinary people. I believe common sense is an automatic, emergent byproduct of a certain type of cognitive architecture, and that connectionist architectures – even in simple stripped-down domains – are much more likely to exhibit common sense than are the fanciest traditional symbolic architectures, whether in knowledge-intensive or in stripped-down domains. I think it is important to clarify what is meant by "common sense" and to strip it of its mythical dimensions.

For purposes of clarification, therefore, let us consider a scenario that clearly calls for commonsense thinking. Suppose you have invited a close friend for dinner, and she doesn't show up. From past experience, you know her to be very reliable. As the hour grows later and later, what do you do? There are all sorts of possibilities, including these fairly obvious and sensible ones: phone her home; go over to her place, if it's not too far; guess her most likely route to your place, and trace that route; phone the police; start eating dinner yourself.

As time passes and you become increasingly concerned, various less obvious but still fairly reasonable steps will come to mind, perhaps including the following: check your own calendar to make sure *you* didn't get the day wrong; phone your friend's neighbors, friends, or relatives, for ideas or clues; phone or go to her favorite haunts and see if she's at any of them; go back to her place and leave a note on her door; scour her yard and perhaps try to get into her house; phone the local hospital; phone her employer at home.

Once these thoughts are exhausted, you begin to get desperate, and therefore some far-out possibilities start coming to mind, such as: start wondering if you actually *did* invite her to dinner, after all; write her an angry letter and tell her she's no longer your friend; call up somebody else to take her place; ask a radio station to broadcast an announcement asking her to contact you; hire a psychic or fortune-teller to help locate her.

Although it would be impossible to draw an exact boundary line, there is a point at which the ideas that come to mind verge on the irrational. In fact, the following are ideas that would occur to a rational person only as humorous thoughts lightening up the serious mood, if they occurred at all. These avenues are exceedingly unrealistic and some of them would require genuine creative intellectual effort to come up with: take a plateful of your dinner and leave it on your friend's porch; engage a pilot to sky-write a note asking her to contact you; turn on the ballgame on TV and scour the bleachers to see if she might have gone there; call the New York Public Library reference desk for help; write to Miss Manners for advice.

This thought experiment conjures up an image of a "sphere of possibilities" centered on the given situation, where distance from the center indicates, very crudely, the degree of implausibility involved. Another way to conceive of distance from the center is in terms of "tension" or "stress," in the sense that one feels an increasing degree of mental discomfort with the suggestions in outer layers. As desperation mounts, however, a kind of "mental temperature" rises, reducing one's reluctance to sample regions of this sphere far from its center. This type of mental temperature can be taken as a measure of *tolerance of tension*.

Let us refer to this sphere of possibilities as the *commonsense halo* surrounding the mental representation of the situation in question (here, that your friend has not shown up). How do human beings access elements of such a commonsense halo in a plausible order?

One conceivable way of exploring such a halo would be to generate all the possible ideas (of which there are a limitless number, of course) and then to rank them in terms of ease of execution, likelihood of success, and so on, somewhat as a game-playing program looks ahead in a full-width move tree. Clearly this is nothing like what people do. People effortlessly generate closer-in, more commonsensical ideas without having to edit them from a host of further-out ideas.

A more psychologically plausible method for generating a commonsense halo would be based on *heuristics*. In such a method, each situation would address an appropriate set of heuristics that, given a temperature, would suggest elements of the halo that have the appropriate degree of wildness. The problem with this is that situations are unique, and consequently the addressing mechanisms would have to be extremely sophisticated.

I believe the most psychologically realistic model for the generation of elements of the commonsense halo is based on the idea that each and every concept in the mind is itself surrounded by a halo of neighboring concepts. Such a *conceptual halo* – elsewhere called an "implicit counterfactual sphere" (Hofstadter 1985) – is very much like the commonsense halo described above, except that rather than surrounding a complex *situation*, it surrounds simply one *concept*, such as the notion of contacting someone, or that of "home." Near the core of the conceptual halo around "contact" are such concepts as "phone," "go see," and "write." Further out might be "dream about," "communicate psychically," and so on. These far-out relatives are accessible only at high temperatures.

I would like to make it clear that such a conceptual halo is distributed and has no precise boundaries. I conceive of it as an inevitable, epiphenomenal outcome of "mental topology" – a vision of concepts as intrinsically distributed, overlapping regions in an abstract space (Hofstadter 1979, pp. 652–56; Hofstadter 1984). (One can of course conceive of concepts in a more brain-related way – for instance, as distributed, overlapping groups of neurons – but that is not necessary for a mental topology to exist.) According to this view, some concepts are very near each other (overlap a great deal), others are vaguely related (overlap slightly), while yet others are widely separated (have no overlap). This is hardly novel – it is merely a way of saying that the mind is associatively structured. Therefore, when *concepts* are properly represented in a model (i.e., as overlapping regions in an abstract space), conceptual halos will automatically be present; no extra apparatus will have to be added to the model.

The mental representation of a *situation* (such as a friend not turning up for dinner) is a compound structure involving a number of simultaneously activated constituent concepts, and the commonsense halo around that situation is, accordingly, an automatic consequence of the existence of conceptual halos around all the activated concepts. In order to construct elements of the commonsense halo, it suffices to probe the various conceptual halos involved, one at a time or several in parallel, adjusting the mental temperature as needed. The degree of tension or implausibility attached to a particular element of the commonsense halo is a function of the distances from the cores of the various conceptual halos probed, and thus, indirectly, of the mental temperature. Note the complete lack of heuristics needed, in this model, to account for common sense.

I would certainly not claim to have captured the full complexity of common sense in this sketch, but the imagery is intended

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

to show the intimate relation between common sense and a connectionistic (or at least associationistic) architecture of mentality. Closely related to my claim about common sense and subsymbolic architectures is the further claim that, as researchers attempt to develop increasingly accurate models of such other critical features of ordinary human mentality as memory retrieval, analogy-making, error-making, and creativity, symbolic architectures will reveal themselves to be increasingly brittle (Holland 1986), while subsymbolic architectures will prove to be increasingly supple (Hofstadter 1979, pp. 570–71). Of course, such issues will not be resolved definitively for a long time. In the meantime, philosophical treatises of clarification such as Smolensky's will serve the vital purpose of affording researchers a perspective from the forest level, rather than from the subforest level.

Some memory, but no mind

Lawrence E. Hunter

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

The connectionists have surely done *something*, but no one seems to be certain quite what. Smolensky claims there is a high probability that they will explain all of cognitive science and provide a uniform "theory from which the multiplicity of conceptual theories can all be seen to emerge." Although connectionism has undoubtedly made a contribution to cognitive science, this claim seems untenable.

The first problems with Smolensky's claims arise in the overly broad definition of "connectionism." The original usage (Feldman & Ballard 1982) is more restricted than Smolensky's; there are earlier network models (cf. Grossberg 1976) that meet the broad definition. Smolensky's claims are perhaps best taken to refer to feedforward networks trained using either simulated annealing or back propagation of error.

These networks and training methods contribute to cognitive science a design for content-addressable memory. First proposed in Luria 1966 (see also Kohonen 1984), a content-addressable memory is one in which the address of a piece of stored information can be determined (by the memory store itself) from a retrieval pattern that is similar to the stored pattern. Such a system is crucial to most theories of cognition. Although connectionist models thus far lack some features desirable in a content-addressable memory,[1] they have advantages over discrimination networks (Feigenbaum 1963) and other serial models.

There are, however, competing theories of content-addressable memory (e.g., Hopfield 1982), and connectionist models' performance is substantially worse than state of the art in other domains (e.g., natural language processing [McClelland & Kawamoto 1988] or expert problem solving [Touretzky & Hinton 1985]). Furthermore, several of Smolensky's general claims seem incorrect.

First, cognitive science other than connectionism is not entirely "constructed of entities which are *symbols*," and should not be called the "symbolic paradigm." Some theories in cognitive science do depend necessarily on symbolic manipulation (e.g., variable binding, a touchstone of symbolic processing[2]). Nevertheless, much of the analysis of cognitive science applies equally well to connectionist and nonconnectionist systems. For example, Smolensky's discussion of semantics and rationality has nothing whatever to do with whether the system involved uses symbols or connections; furthermore, it was arrived at more than half a century ago by Tolman. According to action theory (Tolman 1932), organisms strive to map goals to actions, which produce feedback (relative to the goals) that guides change in future mappings. Tolman claimed that one approaches knowledge of the true state of the world through repeated episodes of such goal pursuit with feedback. Smolensky's "subsymbolic semantics hypothesis" is a restatement of this theory, and I do not see how the validity of the claim depends on representations of the environment being internally coded as connections.

Smolensky labels nonconnectionist cognitive science as "competence" theory, calling to mind Chomsky's move to insulate his theory of language from its incorrect predictions about behavior (Chomsky 1980). Smolensky's label suggests, without substantiation, that traditional cognitive theories likewise make incorrect predictions. Perhaps he means that connectionism will be able to make predictions regarding phenomena about which symbolic models must be neutral. This may be the case, although there are also behavioral phenomena about which connectionism must be silent, for example, the effect of synaptic chemistry on reaction time. Theories of cognition are measured by their breadth and predictiveness; Smolensky did not demonstrate that connectionist theories will be broader or more predictive than more (or less) abstract characterizations.

Second, connectionism is significantly incomplete as a theory of learning. Learning, loosely stated, is the improvement of an organism's ability to achieve its goals on the basis of its experience. Clamping the input and output of the system to a desired state is not what is traditionally meant by experience. Even simple, slow learning is more than just forming associations: It also requires deciding how much to attend to which potential stimuli, characterizing stimuli in an appropriate way, and evaluating the relationship between the stimuli and active goals (Schank et al. 1986). In addition, not all learning is simple or slow: for example, learning from single experiences (DeJong 1983), learning to seize opportunities (Birnbaum 1986), and generating novel explanations (Kass 1986). Content-addressable memories are probably necessary, but certainly not sufficient, to perform these kinds of learning tasks. Simulated annealing and back propagation are programming techniques for generating content-addressable memories; they are not models of learning.

Related to the mistaken idea that connectionist systems are models of learning is the claim that "solving the assignment of blame problem is one of the central accomplishments" of connectionism. Smolensky's note 9 belies this claim by explaining that current connectionist systems assign blame by undirected search through the space of possible assignments (tested by repeating the training examples tens of thousands of times). This technique is neither new nor satisfactory. More important, it assigns blame for error in finding the best match in memory, not for identifying which states of the world or actions of the system led to some goal outcome.

Finally, there is a difference between a theory of a content-addressable memory and a theory of what to put in it. Connectionism provides a theory of *how* information is stored in memory, but not *what* information should be stored. Much valuable research has been done into techniques for selecting which objects, relationships, and characterizations should be computed and then stored in order to best further goal pursuit (e.g., Schank & Abelson 1977; Schank 1982; Hammond 1986). These are (in part) theories of *what* to represent. No connectionist training algorithm has created a network that can relate long sequences of sensations and actions to complex goals nearly as well as existing theories.

Despite vociferous claims like Smolensky's, connectionism's contribution has been modest. Content-addressable memory is important, as is the enthusiasm it has whipped up for the field of cognitive science. Connectionism is not a framework for a general theory of cognition, nor for learning, nor even for representation.

NOTES

1. For example, Baron's description of human associative memory (Baron 1987) includes a "goodness of match" measure for each input, and the ability to recall both the best match and associations to it.

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

2. Despite Smolensky's assertion, using essentially the same semantics as English words is not such a touchstone (cf. Schank & Abelson 1977).

On the obvious treatment of connectionism

Stephen José Hanson

Bell Communications Research, Morristown, N.J. 07960; Cognitive Science Laboratory, Princeton University, Princeton, N.J. 08544

1. The claim. It hardly seems controversial that connectionist models can be formally interpreted as doing statistical inference and minimizing differential equations. The controversial aspect of Smolensky's target article concerns the way he wishes to characterize these familiar, numerically relevant mathematical systems. His claims seem to be motivated by what might be called the "strong implementational view of connectionism."

If a symbolist theory of, say, phonology exists, then showing that it can be represented in a connectionist system provides no new information about phonology or about what the theory of phonology should look like. Connectionist modelers must look at the theory of phonology and use insights from that theory to develop their connectionist model. It becomes *merely* an implementational account.

Given the strong implementational view, it would seem critical to be able to show that connectionist models have some *special* properties that give them new computational abilities and representational semantics that do not or *in principle cannot* appear in symbolist accounts. What is needed, according to Smolensky, is the subconceptual level – with subsymbols and subthoughts and subbehavior and subcognition – which represents the proper level of analysis for the study of the mind and cognition. Smolensky argues that these "subthings" are somewhere between things and neurons; consequently, they are neither things nor neurons but can be made to approximate things and neurons.

Let me try to list some of the presuppositions of this characterization of connectionist models and to provide an alternative account of what connectionist models gain from distributed representation and why connectionist models are not *merely* implementational.

2. Symbolist theories aren't complete, nor are they correct. One presupposition of Smolensky's approach seems to be that symbolist (rule-based) accounts of psychological phenomena are correct, complete, consistent, and served up on a silver platter. I daresay it would not be hard to find lots of counterexamples to this assumption.

A second tacit assumption of this strong implementational view seems to be that theory development is not affected by the medium and axioms of the model used to implement the ideas. Although verbal theories are not that easy to come by, it is just as hard to express theories in a detailed formal system, perhaps harder. Much of the original intent of the theory and goals may be lost in a particular formalization.

It is clear, however, that the constituents and structure of the model can help or perhaps impede theory development; the theory and the modeling environment interact to make the parts of the theory vulnerable and to bring out relations among variables that the theory may only hint at or not refer to at all. Connectionist models may provide just the sort of constituent structure that many symbolist theories badly need.

3. Differential equations and symbols can get along. Do we really need a new language and terminology for standard mathematical systems and their effects? Differential equations have had a long history in the natural sciences and they of course differ from recursive rule systems; probability models likewise differ from boolean models in their form of expression. What becomes difficult to reconcile is (1) the technical jargon that arises within the mathematical system and (2) folk descriptions

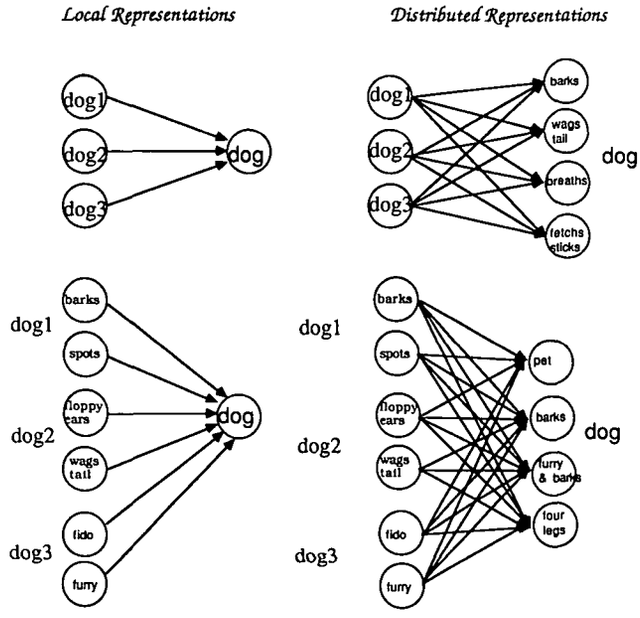

Figure 1 (Hanson). The four possible kinds of representational relations between tokens and types. On the left are two kinds of hidden-layer localist representations. On the right are two kinds of distributed representation.

of the phenomena, which are approximate, inaccurate, and intuitive. Connectionist jargon has already begun to include intuitive notions of optimal points, local minima, gradient search, and general optimization notions. This language will probably evolve naturally with the mathematical theory of networks and their relation to the phenomena they model. At this juncture the notion of "subthings" may obscure rather than clarify the distinctions between symbolic and connectionistic modeling.

4. Subsymbols or features? Let me be precise about a simpler alternative to Smolensky's PTC (proper treatment of connectionism); let me call it OTC (the obvious treatment of connectionism). In Figure 1, I have displayed the possible kinds of network configurations as a function of input or output unit representation and the hidden unit representation adopted through learning (local or distributed). The first case shown in Figure 1 is what we might associate with the usual symbolic or rule-based approach; it is of course quite "localist" (i.e., a single unit constitutes a single symbol). Here the tokens "dog1," "dog2," and "dog3" are being mapped to the "dog" concept. This type of process is the assignment of a set of tokens to a type. It is up to other operations to associate the tokens with other tokens or featural representations of the same type as in, for example, an inheritance process.

The second case shown is still a kind of symbolist representation as well as a case of local representation. In this case, however, tokens are first decomposed into a set of primitive features or composite types which cover all the possible tokens in the domain. So in the present case, "barks," "has fur," "breathes," and "has spots" are examples of a set of features that might be used to describe dogs and other animals. Schank's (1975) Conceptual Dependency approach is an example, with the tokens in the domain first mapped onto a set of general types or features (P-TRANS, a physical transfer of information). Other operations such as planning or problem solving would involve the manipulation of this sort of general feature information.

The third case is the first simple instance of a distributed representation. The hidden layer is representing the tokens as a set of types or features that would cover all the possible tokens in the domain. In fact, exactly the same kinds of features could be used as in the second localist case just mentioned. The only

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

difference between this distributed representation and the localist representation is that the feature decomposition is a function of the weight connections which encode the relations between the tokens "dog1" (etc.) and the feature values in the hidden layer representing the "dog" concept, "barks" and so on.

The fourth and last case is completely distributed between the input representation and the hidden layer. In this case there exists a featural representation in both the input layer and the hidden layer. It would be most useful to have specific features with specific tokens (dog1 is composed of "fido," "barks," "floppy ears," etc.) and to allow these to be recombined in the hidden layer in order to "construct" the dog concept. We might also allow new features to be constructed or to emerge from a "decomposition" of our originally chosen features. So in the hidden layer "dog" tokens would tend to activate "bark," "breaths," "wags tail" and perhaps co-occurrences (or n-tuples) of general features, "barks and wags tail." In this case we would not have to indicate which hidden unit belonged to which feature in the input layer. These hidden-unit feature-bindings could be discovered during learning.

At this point one might wonder what is so remarkable about distributed representation. The four cases we have discussed are not really antithetical to symbolist approaches, nor do they provide any exceptional new view on representation that symbolists haven't already thoroughly considered. Featural representations are not new or very difficult to characterize; they date back to Aristotle.

What then distinguishes the connectionist representations from any other kind? We begin to see a bit of it in case 3 where tokens are being mapped to featural representations. The difference between this case and case 2, where featural representations are chosen a priori for the domain and its tokens, is that the representation between the token and its features is made visible. It is now obvious from the connection strengths what token "dog1" is made up of, plus the representations that determine its composition are readily accessible, visible, and shared[1] among all other tokens in the domain. This visibility or accessibility of the network representation is what distinguishes it from the localist/symbolist representations in which such information has been committed to the representation language. This difference is one of the reasons why *learning* is possible and so natural for connectionist networks. In contrast, symbolic approaches must somehow make clear to the learning operations they use what aspects of the representation were responsible for some event; and worse, they must track down the "hidden" information initially assumed or even axiomatized in the representational language and make it visible to those same learning operations.

This focus on learning now makes it crucial to understand the intentional feature and mapping operations once they have been learned. In some sense the representation problem has been turned on its head. Instead of asking what the proper feature representation is for a cup or a chair, connectionists want to know under what conditions the proper featural representation would be learned for cups or chairs.

NOTE

1. This leads to the problem of how to extend the feature "lexicon" of the network. This too is not new. Any representational scheme that uses features must encounter this constraint at some time or another.

Smolensky, semantics, and the sensorimotor system

George Lakoff

Linguistics Department and Institute for Cognitive Studies, University of California, Berkeley, Calif. 94720

I admire Smolensky's attempt to characterize the relationship between connectionist research and more traditional issues in cognitive science. My comments are of two sorts: some clarifications where I think Smolensky might have said things a little better, and some important areas that he did not treat, but which are consistent with his overall approach.

Some clarifications. In mentioning the "conceptual level", Smolensky does *not* mean to return to the symbol-manipulation paradigm. His "conceptual level" is not a kind of logical form (say, of the old generative semantics variety) nor a Fodorian "language of thought." Smolensky's "conceptual level," as I understand it, would have to conform to the mathematics of dynamical systems, and not to the mathematics of recursive function theory and model theory. I will say more below on how that might be done.

A possible misunderstanding may arise from Smolensky's use of the word "level." This word is used in the academic world in at least two senses. In linguistics, levels are taken to be distinct representations of different kinds, with correspondences between elements across levels. For example, many linguists speak of the phonetic level, the syntactic level, and the semantic level, with the assumption that these are three different kinds of representations. This is not what Smolensky has in mind. Instead, he has in mind something more like the physicists' notion of level, as in the subatomic level, the atomic level, the molecular level, and so on.

Thus, Smolensky's three levels are not three different kinds of things. There is the neural network of the physical brain: This is the neural level. There is the aspect of the physical brain (namely, the neural structure and activity) that connectionism picks out to model: This is the subconceptual level. And there is a structure to the activation patterns of that aspect of the brain's neural network that connectionism models: This is the conceptual level. Both the subconceptual and conceptual levels are aspects of the neural networks of the physical brain and their activity.

I assume that this is what Smolensky has in mind, and will proceed from here.

Semantics and the sensorimotor system. Smolensky's discussion makes what I consider a huge omission: the body. The neural networks in the brain do not exist in isolation; they are connected to the sensorimotor system. For example, the neurons in a topographic map of the retina are not just firing in isolation for the hell of it. They are firing in response to retinal input, which is in turn dependent on what is in front of one's eyes. An activation pattern in the topographic map of the retina is therefore not merely a meaningless mathematical object in some dynamical system; it is *meaningful*. A different activation pattern over those neurons would mean something different. One cannot just arbitrarily assign meaning to activation patterns over neural networks that are connected to the sensorimotor system. The nature of the hookup to the body will make such an activation pattern meaningful and play a role in fixing its meaning.

Compare this, for example, with a string of symbols in a Fodorian language of thought, or in a computer program. The symbols are not meaningful in themselves. They have to be "given meaning" by being associated with things in the world. If the symbols are to stand for categories, those symbols must be given meanings by being associated with categories that are out there in the world. In my recent book (Lakoff 1987) I survey a wide range of evidence showing that such a project is impossible, that the symbolic paradigm cannot have a viable theory of meaning.

Interestingly enough, in the evidence I survey there is not evidence against a connectionist account of meaning. The reason is that activation patterns over neurons can be meaningful in themselves when the neurons are appropriately located relative to the sensorimotor system. Such activation patterns do not have to be "given meaning" the way that strings of symbols do.

The point of all this is that, counter to what critics like Fodor, Pylyshyn, Pinker, and Prince have said, it is connectionism, not

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

the symbolic paradigm, that is the only game in town. And it is the connection to the body that makes connectionism a player in the semantics game.

Invariance and cognitive semantics. Connectionist semantics is, of course, not highly developed at present to say the least. But all that could change in a short time. The reason is that cognitive semantics, as it is being developed within linguistics, meshes well with connectionism.

The basic mechanisms of cognitive semantics include cognitive topology, mental spaces, metaphor, and metonymy. Technically, cognitive semantics is consistent with the connectionist paradigm but not with the symbol-manipulation paradigm. One reason is that cognitive topology, which provides for the basic mechanisms of reasoning, is continuous rather than discrete.

At present there is a gap between connectionism and cognitive semantics: We do not know how cognitive topology can be implemented in connectionist networks. Such an implementation should be possible. The key, I believe, is what I have been calling the "invariance hypothesis." The idea is this: Each of the elementary structures of cognitive topology – bounded regions, paths, contact-versus-noncontact, center-versus-periphery, etc. – have to be preserved in mappings from one modality to another in order for sensorimotor control to be successful. It is hypothesized that activation patterns corresponding to such structures arise in the development of sensorimotor control, and are mapped onto structures[1] of abstract reason by the connectionist mechanism for characterizing metaphor: mappings from one neural ensemble to another across a narrow channel. The structures studied in cognitive topology are called "image schemas" (or sometimes merely "images"). The best places to read about them are in Lakoff, 1987, case study 2 and in Langacker, 1987.

Conclusion. In applying connectionism to issues in cognitive science, it is important not to think of it as just another mode of information processing, in parallel instead of in sequence. In a full-blown connectionist theory of mind, activation patterns over neurons are meaningful in themselves by virtue of what those neurons are connected to. The intractable problem of assigning meanings to symbols does not arise here.

It is also important to remember that the isolated models connectionists build to study the properties of networks are not full-blown connectionist theories of mind. They vastly oversimplify, or totally ignore, sensorimotor input and output, assuming that, for the purpose of the study at hand, one can just as well use feature names, to which the model-builders must assign meanings. This is a crucial difference between isolated models and a full-blown theory.

There is another important difference. What neural networks can do is constrained in the full-blown theory by the nature of the sensorimotor system. For example, consider what connectionist phonology in the full-blown theory would be like. Phonological processes, in large measure, would be characterized by conventionalized activation patterns controlling articulatory and acoustic processing. This would help to limit the general principles embodied in phonological patterns to those that are phonetically realistic. In symbol-manipulation phonology (that is, generative phonology), no such restrictions are automatically built into the theory. However, since such sensorimotor constraints are not built into the isolated models, those models do not embody the constraints of the full-blown theory. Thus, where the full-blown theory can offer phonetic explanations for constraints on phonology, the isolated models cannot.

For such reasons, it is vital to bear in mind that a full-blown connectionist theory of mind is a lot more than just an information-processing system.

NOTE

1. For example, idealized cognitive models, grammatical constructions, image-schemas, etc. (see Lakoff 1987 for details).

Physics, cognition, and connectionism: An interdisciplinary alchemy

Wendy G. Lehnert

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

As the symbolic/subsymbolic debate rages on I've noticed many of my colleagues in the so-called mainstream symbolic AI community backing off from public debates on the matter. In AI, one normally doesn't talk about anything for more than two years unless the idea is generating about a dozen Ph.D. theses. But that's just one part of the story. In truth, a lot of us have first-hand experience with graceless degradation and we understand very well about the desirability of soft constraints: My semantically oriented colleagues in natural language have understood about such things ever since the early days of preference semantics (Wilks 1978). Even so, the desirability of soft constraints does not negate the validity of the symbolic paradigm. Although Smolensky does not advocate that we abandon symbolic information processing, there is nevertheless something facile about his conciliatory conclusion depicting one big happy family where everyone can peacefully coexist.

The first thing I noticed about connectionism was how the psychologists picked up on it long before the AI community got interested. Initially this seemed puzzling to me, but then it made perfect sense. The interdisciplinary appeal of connectionism is not so much a computational appeal, as it is an appeal based on theorem envy. We must also understand that the problem of theorem envy has always been stronger in psychology than it ever was in AI. This is undoubtedly because graduate students in computer science who harbor a strong desire to be mathematicians have the option of becoming theorists in computer science. In graduate psychology programs, there seems to be no analogous safety valve for those seeking rigor in their lives. Or at least there wasn't until the early '80s.

In recent years connectionism has come to the rescue of a new generation of psychologists who are really closet mathematicians and physicists. Unfortunately, there is one aspect of theorem envy which is a serious threat to the health of cognitive science: methodology-driven research at the expense of problem-driven research.

Here is where I find connectionism potentially dangerous: Most connectionists are methodology-driven. The connectionists who claim to be doing neural modelling are clearly methodology-driven (see, for example, Churchland 1986). Even the researchers who distance themselves from neural modelling are methodology-driven in slightly more subtle ways. Smolensky is a good example of this. A central thesis of his target article places subsymbolic processing above the neural level, so Smolensky's view of connectionism does not derive from neurophysiology. Rather, Smolensky hopes to wed the natural sciences to cognition by deriving the cognitive principles underlying subconceptual processing from physics. He takes a narrow view of dynamical systems as the proper foundation for subsymbolic processing, and then attempts to distinguish dynamical systems that are cognitive from dynamical systems that are purely physical. This is a methodology-driven argument which positions the methodologies of physics at the center of Smolensky's view of connectionism. However powerful these trappings are for those who feel reassured by equations, a fondness for physics and its associated mathematics has narrowed Smolensky's view of connectionism, cognition, and computation in general.

We see Smolensky-the-physicist at work in the "connectionist dynamical system hypothesis" which describes an intuitive processor in terms of differential equations. Not content to stop there, he proposes a definition for cognitive systems which implies that the only difference between a cognitive person and a noncognitive thermostat is a matter of degree (pardon the

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

pun). Putting aside the problem of defining cognition on a slippery slope, there are aspects of Smolensky's perspective that seem fundamentally correct to me. For example, I agree with his emphasis on process over structure. But we should remember that this shift in perspective was one of the foundational contributions of symbolic AI.

While it is indeed wonderful to see a connectionist system exhibit nonmonotonic inference with respect to the past-tense verb production (Rumelhart & McClelland 1986), let us not forget that one of the most exciting things about EPAM was its ability to exhibit oscillation effects as it learned (Feigenbaum 1963). There is a remarkable tendency for methodologically driven researchers to feel justified in dismissing as inconsequential any work done outside of their own methodology.

A problem-driven researcher is happy to use any techniques or methodologies that fit the problem at hand. If a computer program can be described in terms of differential equations, one should not hesitate to exploit appropriate mathematical results. Bravo. But Smolensky is not content to stop there. He wants to argue that the connectionist doorway into nondiscrete mathematics presents profound and revolutionary implications for the study of cognition. This is where Smolensky and I must part company.

The role of discrete versus continuous mathematics in theories of cognition is a nonissue if we resist methodology-driven reasoning. The more important question is whether or not knowledge can be modularized, accessed at variable levels of abstraction, and manipulated with procedures appropriate to those different levels of representation. That is, the real problems we should address in trying to distinguish subsymbolic and symbolic processing are representational problems.

Unfortunately, knowledge representation is not one of the strong points within the connectionist paradigm. To say that knowledge in a connectionist computer program is manifest within a system of weights and differential equations is not only a retreat from symbolic meaning representation, it's a retreat from meaning altogether. In fact, I would argue that the concept of "distributed representations" works for connectionists the same way the concept of a "kludge" works for symbolic AI researchers. They are both absolutely necessary, undeniably convenient, abused with wild abandon, and potentially disastrous in the long run. If pressed, we throw up our hands and admit that we don't quite understand what we're doing here. But a clever kludge is not immediately apparent – and distributed representations are quite clever: It's very hard to see what's wrong at first glance.

The distributed view of representation promises to deliver a lot of tempting goodies: Soft constraints, graceful degradation, "best fit" pattern matching algorithms, learning from examples, and automatic generalization are nothing to sneeze at. Unfortunately, the reductionistic nature of a distributed representation also makes it extremely difficult to do the simplest things. We can identify a room based on a description of its furniture (Rumelhart, Smolensky, McClelland & Hinton 1986), but we have no natural way of identifying the relationship between the room and the furniture – they are only associated with one another in some unspecified amorphous manner (Charniak 1987). A similarly fundamental problem arises with variable bindings (Feldman & Ballard 1982; Shastri & Feldman 1985).

To be fair, a lot of connectionists are seriously addressing the question of representational power (Hinton 1986; Cottrell 1987; McClelland 1987; Pollack 1987; Shastri 1987; Touretzky & Geva 1987), so we cannot assume that Smolensky has chosen to ignore representational issues for lack of activity in this area. Rather, Smolensky appears to acknowledge the role of representational power only as a "conceptual" issue which divorces it from the concerns of subsymbolic processing in his dichotomy.

If we accept Smolensky's criteria for separating the conceptual from the intuitive and the symbolic from the subsymbolic, it is easy to go along with his view of "soft adaptive" processes underlying "hard and brittle" rule application. If representational issues are defined to be purely conceptual phenomena which need never intrude into the formal sanctity of numeric vectors and differential equations, then the prospect of symbolic/subsymbolic turf wars does indeed seem remote. But I'm not buying the Smolensky scenario. Representation is closer to the heart of the matter than Smolensky would have us believe.

Can this treatment raise the dead?

Robert K. Lindsay

Mental Health Research Institute, University of Michigan, Ann Arbor, Mich. 48109

Much of the appeal of connectionism is that it is a form of associationism with a long history in philosophy and psychology; associationism's most recent preconnectionist incarnation was behaviorism, whose demise was in large part due to its failure to solve, or even recognize, several key problems for which symbolic models have well-understood and indeed almost obvious natural solutions. Here are the most important: How is it possible to add new knowledge and abilities without disrupting the old? How is it possible to add new knowledge so that it builds on the old? How can knowledge and process be structured hierarchically? How can alternatives be formulated and considered systematically? How can directed, logical, precise thinking be achieved, as it undoubtedly is by humans at least occasionally. The first major question to address to connectionism is whether it can supply solutions to these problems. So far it has not. What does it offer instead?

I think the target article offers two different (though not incompatible) visions of connectionism, and each suggests a different advantage over symbolic models. The first seeks finer-grained and more "accurate" accounts of the macrophenomena allegedly only approximated by symbolic models. In this vision, subsymbolic is to symbolic as quantum mechanics is to classical physics.

The second vision is also reductionist, but it sees the subsymbolic as offering a different *kind* of account. In this vision, subsymbolic is to symbolic as dynamical systems theory is to classical automata theory, or as evolution is to learning theory. These two visions correspond to the two major distinctions between the paradigms and their mechanisms of semantic representation and learning.

To avoid the perennial problem of arguing for substantive differences in computational power in the face of Turing equivalence, Smolensky attempts to distinguish syntactic equivalence from semantic equivalence. If we buy that, then the substantive subsymbolic/symbolic distinction is that symbolic models have single symbols that refer to single concepts, whereas in subsymbolic models the analog of a symbol is a "pattern of activity" among a set of units. *This critical distinction is not made precise in the target article,* and possibly it cannot be.

Indeed, symbolic models often deal with concepts that are not represented in a simple one-symbol-to-one-concept manner. A symbol may denote an internal state of arbitrary complexity and one that changes over time. What distinguishes this representational mode from the subsymbolic one is that it need not be uniform, it is hence potentially richer, and it has a name that other structures can refer to (a crucial advantage).

Smolensky's analysis goes wrong not in describing the subsymbolic, but in an impoverished view of the symbolic level. His characterization of the latter (the most explicit being (24)) is closer to a limited type of symbolic model, rule-based system. If we are denied the refuge of Turing equivalence on the grounds of semantic distinctions, then I reserve the right to distinguish symbolic models in general from the special case of rule-based models in spite of the proven (syntactic) computational univer-

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

sality of production rules. Perhaps recognizing this objection, having given his characterization of symbolism, Smolensky waffles: "any one of them [the discrete characteristics of symbolic models] can be softened, but only by explicitly building in machinery to do so." Of course to a "symbolist," explicitly building in machinery is how the game is played, so to admit that it can be done is a serious concession. Indeed, if one looks at the extant AI symbolic models more broadly they do not uniformly differ from subsymbolic ones in the ways described, but only in explicitly eschewing the desire to be reductions-to-associationism and one-principle learning.

A major case in point is constraint satisfaction. Connectionist writings are at their most compelling when they argue that much of human cognition must be viewed as multiple, simultaneous constraint satisfaction. Many symbolic models, say, of natural language understanding, do indeed prescribe a different view: a serial narrowing of possibilties by a sequence of necessary-condition filters in the form of syntactic and semantic well-formedness rules. However, the critical distinction between this and the constraint view is not whether all constraints can be uniformly represented (they need not and probably should not be), but whether the "rules" of agreement, deixis, logic, physical plausibility, conversational convention, and so on, are viewed as *necessary* conditions; the symbolic paradigm does not require that they should be. Furthermore, these linguistic, logical, and empirical constraints are quite naturally characterized at the symbolic level as a heterogeneous set, and the compelling but informal connectionist arguments for constraint-based processing are usually presented in similar terms; it is unjustified to identify this characterization with the infinitesimal, nonreferential, amorphous "constraints" of subsymbolic models.

The vision of connectionism as reformulation has a literature essentially distinct from the fine-grain-semantics vision just discussed. It is typified by Smolensky's harmony theory, for example, and it attempts to demonstrate how general features of intelligence can arise from a uniform morass of associations by a unitary learning mechanism. The higher cognitive functions are explained indirectly and obliquely by showing how they emerge from a process of adaptation. These efforts have thus far had very limited success.

What Smolensky is offering in Sections 8 and 9 is something quite different from the symbiotic account he offers earlier. It goes beyond an alternative account of established explanations to an entirely new analysis that will raise new questions and supply new answers. For example, one can envision connectionist accounts of the limits of predictability (perhaps based on the concept of "chaos") that explain the landscape of cognition without offering detailed predictions of why John Doe chose chocolate rather than vanilla today, or even why some chess players are better than others.

Connectionism gives up a lot when it abandons the successes of symbolic modelling; ultimately it must replace what is lost, but the path to that integration is not clear. Some of what it hopes to gain – the neuroscience connection – is illusory. Some reputed gains, for example, multiple soft constraint satisfaction, are well within the symbolic paradigm in principle if not in current fashion. One claimed accomplishment – a general learning mechanism – is thus far unproven on problems of realistic magnitude, and the suggested quest for parsimony is probably premature. But what remains could be the seed of a reformulation of the goals and scope of cognitive science.

Connectionism in the golden age of cognitive science

Dan Lloyd

Department of Philosophy, Trinity College, Hartford, Conn. 06106

Subsymbolic and symbolic paradigms. Smolensky rejects hypothesis 10, the ecumenical view that connectionist models are mere implementations of symbolic models and their conceptual-level explanations, a view implying that orthodox cognitive science (the Newell-Simon-Fodor-Pylyshyn view) remains the only genuinely cognitive game in town. Against this, Smolensky asserts (8c) that "complete, formal, and precise descriptions" of intuitive cognition will only emerge at the subconceptual level of connectionist models. (The closing lines of Section 2 seem to extend this claim to cognition in general.) Apparently the argument is roughly this: Since there will be no neat formalisms at the conceptual/symbolic level, there is nothing for the neat formalisms at the subconceptual/subsymbolic level to implement. This is a weak ground for the autonomy of connectionism: Who knows whether there may someday emerge a neat conceptual level model of cognition? Properly treated connectionism should not, I think, stake its future on the *failure* – a scruffiness that may be corrigible – of an allied cognitive enterprise.

Rather, connectionism might well illuminate the successful formal treatment of cognition at the conceptual level. The autonomy of connectionism would then rest on distinctions *between models* (at the conceptual level), rather than on the distinction *between levels*. Models at the conceptual level should be the central aim of connectionism in any case: After all, we seek true psychology rather than subpsychology; we want our *neatest* formalizations to quantify over thoughts, beliefs, and representations in general, rather than over merely *sub*-thoughts, subbeliefs, and subrepresentations.

The details of the conceptual level connectionist model of cognition are presently an open question, perhaps *the* open question, of connectionism. The familiar connectionist choice between local and distributed representation understates the theoretical challenge of conceptual level connectionism. First, "distributed representation" remains ambiguous in connectionist parlance. Sometimes individual processing units are assigned to *features* in the task domain, as, for example, in the circuit analyzer model described in Section 9.2. But sometimes, in contrast to featural representation, individual units are beneath all interpretation. Such representations are *fully* distributed, and no unit is dedicated to the representation of any particular aspect of the task domain. But discussions of connectionism rarely distinguish these interpretative options.

More important, in practice most connectionist models use a mix of representational styles, often with local or featural representations sandwiched in between. As Terry Sejnowski (personal communication) and Smolensky (Section 3) both observe, this heterogeneity makes the interpretation of networks *at the conceptual level* extremely complex. Expressed at the conceptual level, representations are local (since conceptual-level hypotheses refer to representations rather than their substrate), but their dynamics are complex, and *not* the dynamics of subconceptual unit interaction (expressed by the activation and learning equations). New interpretative applications of the analytic tools of linear algebra will be welcome here. (Smolensky, 1986 and Section 3, and Sejnowski and Rosenberg, 1987, are among the pioneers in this enterprise.) I think it's in keeping with the spirit of the target article and connectionism overall to aspire to discover neat formal principles at this higher level, since the principles will be distinctly connectionist.

Connectionism and neural models. Smolensky also distinguishes connectionist models from neural models. Hypothesis (12) summarizes the differences; (12c) reminds us that we don't know much about the neural details and seems to be the

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

premise on which the distinction rests. But our ignorance of the neural details is another *temporary* problem. Like the distinction between connectionist and conceptual-level hypotheses, this distinction is a shifty base for the autonomy of connectionism. Suppose we had all the relevant neural details, and used them to build a model of cognition. Neural dynamics are more complex and heterogeneous than those of current connectionism, but I presume that it isn't essential to connectionism that its dynamic principles be as simple as those of today's models. A total *neural* model would therefore nonetheless be a connectionist model.

Connectionism also needs all the neural inspiration it can get. At this writing, about one year has passed since the thundering arrival of the two-volume cavalry charge of the San Diego PDP group (Rumelhart & McClelland 1986; McClelland & Rumelhart 1986). Workers loyal to the symbolic paradigm have had a year to rally their forces, and as these words see print the cognitivist countercharge should be underway, with each connectionist model the ground of a pitched battle. I expect the pennant in 1988 to be a toss-up: Both the subsymbolic and symbolic approaches will celebrate their successes, and each will be able to point to flaws and omissions, often in great detail, of the other approach. But there is one important foundation for cognitive modelling exclusive to connectionism, and that, of course, is its "neural inspiration." At present, it would be bad tactics to abandon the goal of incorporating as much neural reality as possible. More important, the likelihood of connectionist models being *true* increases with the incorporation of neural dynamics.

The golden age. Smolensky's hypothesis 11 posits a fundamental level for the subsymbolic paradigm, distinct from both the conceptual and neural levels. Following from the discussion above, I suggest the following substitution for hypothesis 11:

> (11) The fundamental level of the subsymbolic paradigm encompasses *both* the neural and conceptual levels.

The golden age of cognitive science will be one in which (thoroughly understood) neuroscience (thoroughly) informs (thoroughly understood) conceptual level cognitive psychology. Connectionism as an autonomous science in the middle serves to catalyze the development of the golden theory, but one upshot of the two compressed discussions above is that connectionism, as a discipline in the middle characterized by the straightforward dynamics of numerous homogeneous processors, will fade away. Its sublimation is no loss, however, since the science it establishes will be connectionist in spirit. It will be the fulfillment, not the refutation, of the promising approach exemplified in Smolensky's target article.

Symbols, subsymbols, neurons

William G. Lycan

Department of Philosophy, University of North Carolina, Chapel Hill, N.C. 27514

This decade past, the philosophy of cognitive science has mongered a number of closely related distinctions: Software vs. hardware; dry abstract computation over predicate-calculus formulas vs. wet biologic cell chemistry; printed circuitry vs. warm fuzzy squirmy animals; pleasant air-conditioned high-tech computer center vs. cleaning out smelly cages; MIT vs. southern California. Great virtues of Smolensky's target article are his rejection of such stereotypes, his recognition of the lush multiplicity of levels of nature, and his attempt to clarify the relation between several of the levels as they are simulated in connectionist computer programs. Pardon my mentioning it again (Lycan 1981; 1987), but the all-too-common two-Level picture of nature – of brains, in particular, or even of computers themselves – is both completely untenable and responsible for many very bad ideas in the philosophy of psychology.

I want to address the compatibility issue raised by Smolensky in his Section 2.4, particularly since it bears on the methodological advice he offers in closing. According to the "subsymbolic paradigm," only the subconceptual level, not the conceptual level, affords "complete, formal, and precise descriptions of the intuitive processor." Ipso facto, Smolensky argues, the "symbolic paradigm" would be ruled out, since it claims precisely to afford such descriptions at the conceptual level. Thus incompatibility.

That seems right on its face, but we should consider a simple irenic response: In his next section (see also 7.1), Smolensky grants that *representation* occurs within (not just as an epiphenomenon of) the "subconceptual" level, and that connectionist models key on "fine-grained features such as 'roundedness preceded by frontalness and followed by backness'" (sect. 3, para. 2). The obvious objection is: Why does this keying itself not count as fully conceptual, fully symbolic activity? A first reply might be that the features or concepts thus mobilized are not ones that occur in the subject's own working vocabulary. But nothing in the symbolic paradigm implies that they should.[1] A second reply might be that the topology is all wrong; symbolic-paradigm computation is linear, prooflike, discrete, monotonic, and so on, whereas connectionist architecture differs in the fairly drastic ways Smolensky has described. But if subsymbols are still representors manipulated according to precise rules, they are still symbols, expressing concepts, in any traditional sense of those terms;[2] the only question remaining concerns what the rules actually are. So far, the subsymbolic paradigm seems to belie its own claim of relocating real cognition to a truly *sub*conceptual level of description (and so it threatens its own alleged claim [8c]).

There is of course the issue of morphology. In the symbolic paradigm, a representation-*token* is a fairly salient chunk or stretch of hardware-at-a-time. In the subsymbolic paradigm, the token – though it exists – is distributed or highly scattered through the system, morphologically foggy or invisible. That paradigm difference is potentially important to computer science and to psychology, for all the reasons Smolensky presents. But its importance to the theory of representation generally is less clear. I can think of just three differences it would make in turn: (A) Though the subsymbolic paradigm allows for higher-level intentional reference to the external world by regions of hardware however scattered, such reference could not (according to the paradigm) be characterized in the terms normally considered appropriate to the computationally relevant higher level. As Dennett (1986, p. 69) has noted, the "brain-thingamabob [that] refers to Chicago" would *per se* have to be described statistically and in terms of the whole connectionist system or a very large mass of it. (B) The subsymbolic paradigm vastly complicates any account of the intentionality of ordinary folk-psychological representations and of anything in standard "symbolic" theory that is at all like them. In Section 7.1 Smolensky makes a start at trying to capture ordinary mental reference, but his attempt is both vague in the extreme and apparently circular. (C) As Smolensky argues in Section 9.2, the subsymbolic paradigm sheds considerable light on the vexed competence/performance distinction, as the symbolic paradigm does not. And I would add that it makes that distinction cognate with Davidson's (1970) otherwise troublesome thesis of the "anomalism of the mental"; if the subsymbolic paradigm is right, then there are no strict psychological laws that can be couched in commonsensical English or even in real-time linear-computational terms. On the other hand, we can profitably see such laws as true under natural if extreme idealizations of "well-posedness" and unlimited time.

To settle the incompatibility issue we would have to bicker

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

further over just what is required for being genuinely "conceptual" or "symbolic." Questions (A)–(C) are perhaps more important for cognitive science and its metatheory.

NOTES

1. Except possibly the idea that the symbols in use at the conceptual level are closely derived from those occurring in the propositional attitudes posited by folk psychology. But that idea plays no role in the issue that officially concerns Smolensky.

2. Barring one in which conceptual activity must be *conscious*. But here again, despite Smolensky's occasional allusions to consciousness, the notion has no relevance to the debate between his two paradigms.

Epistemological challenges for connectionism

John McCarthy

Computer Science Department, Stanford University, Stanford, Calif. 94305

1. The notion that there is a subsymbolic level of cognition between a symbolic level and the neural level is plausible enough to be worth exploring. Even more worth exploring is Smolensky's further conjecture that the symbolic level is not self-sufficient, especially where intuition plays an important role, and that the causes of some symbolic events must be explained at some subsymbolic level. That present-day connectionism might model this subsymbolic level is also worth exploring, but I find it somewhat implausible.

An example of Smolensky's proposal is that the content of some new idea may be interpretable symbolically, but how it came to be thought of may require a subsymbolic explanation. A further conjecture, not explicit in the target article, is that an AI system capable of coming up with new ideas may require a subsymbolic level. My own work explores the contrary conjecture – that even creativity is programmable at the symbolic level. Smolensky doesn't argue for the connectionist conjectures in his paper, and I won't argue for the logic version of the "physical symbol system hypothesis" in my commentary. I'll merely state some aspects of it.

2. The target article looks at the symbolic level from a certain distance that does not make certain distinctions – most important being the distinction between programs and propositions and the different varieties of proposition.

3. My challenges to connectionism concern epistemology only – not heuristics. Thus I will be concerned with what the system finally learns – not how it learns it. In particular, I will be concerned with what I call *elaboration tolerance*, the ability of a representation to be elaborated to take additional phenomena into account.

From this point of view, the connectionist examples I have seen suffer from what might be called the unary or even propositional fixation of 1950s pattern recognition. The basic predicates are all unary and are even applied to a fixed object, and a concept is a propositional function of these predicates. The room classification problem solved by Rumelhart, Smolensky, McClelland and Hinton (1986) is based on unary predicates about rooms, e.g. whether a room contains a stove. However, suppose we would like the system to learn that the butler's pantry is the room between the kitchen and the dining room or that a small room adjoining only a bedroom and without windows is a closet. As far as I can see the Rumelhart et al. system is not "elaboration tolerant" in this direction, because its inputs are all unary predicates about single rooms. To handle the butler's pantry, one might have to build an entirely different connectionist network, with the Rumelhart et al. network having no salvage value. My epistemological concerns might be satisfied by an explanation of what the inputs and outputs would be for a connectionist network that could identify all the rooms of a house, including those whose identification depends on their relation to other rooms.

I might remark that the 1960s vision projects at Stanford and M.I.T. were partly motivated by a desire to get away from the unary bias of the 1950s. The slogan was "description, not mere discrimination." Indeed, one of the motivations for beginning to do robotics was to illustrate and explore the fact that to pick up a connecting rod a robot needs to do more than just to identify the scene as containing a connecting rod; it requires a description of the rod and its location and orientation. Perhaps connectionist models can do this; and it seems to me very likely that it can be done subsymbolically. I hope that Smolensky will address this question in his response to the commentaries.

A semi-heuristic question of elaboration tolerance arises in connection with NETTALK, described by Sejnowski and Rosenberg (1987). After considerable training, the network adjusts its 20,000 weights to translate written English into speech. One might suppose that a human's ability to speak is similarly represented by a large number of synaptic strengths learned over years. However, an English-speaking human can be told that in the roman alphabet transcription of Chinese adopted in the People's Republic of China the letter Q stands for the sound |ch|, and the letter X for the sound |sh|. He can immediately use this fact in reading aloud an English text with Chinese proper names. Clearly this isn't accomplished by instantly adjusting thousands of synaptic connections. It would be interesting to know the proper connectionist treatment of how to make systems like NETTALK elaboration tolerant in this way.

In defence of neurons

Chris Mortensen

Department of Philosophy, University of Adelaide, S.A. 5001, Australia

I take up Smolensky's proposition (17) with which I am in agreement. J.J. C. Smart (1959) suggested nearly thirty years ago that there are various conscious judgements we make about ourselves which have fairly direct neural correlates. We judge a state (a mental state) of ourselves to be waxing or waning for example; and it is reasonable to think that some activity really is waxing or waning. Clearly the activity being dealt with will be a waxing and waning in the relatively large-scale spatial and temporal structure of activity patterns. Smart also claimed to account for conscious judgements about perceptual states, such as colour experiences, as awarenesses of characteristic patterns of *similarities and dissimilarities* between broad structures of neural processes, without anything further being present to consciousness about the features of those structures responsible for the similarities. The point is that the aspects of waxing and waning are plausibly understood in terms of changes in gross summation or averaging of levels of individual neuronal activity, admittedly of distributed patterns of activities identified to consciousness in other ways as well. So some aspects of "conscious phenomenology" are fairly close to the neural level (I mean the level of neural *concepts*, not of individual neurons).

The idea that consciousness is a relatively coarse-grained register of neural activity works less well for the phenomenology of some of our perceptual states, however. Smolensky's subsymbolic methodology accounts well for undefinable judgements of similarity or rightness which are part of the intuitions of the experts, as it does for "the loss of conscious phenomenology with expertise." It is nonetheless less plausible to claim with Smart that the contents of consciousness are *mere* similarities and dissimilarities in the case of, say, colours. The phenomenology of colours remains intractable, I would say.

Smolensky allows that visual and spatial tasks might be an area where the subsymbolic and neural levels merge, a view which I

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

would argue is supported by evolutionary considerations. It is not such an unreasonable speculation that information about shape and spatial relationships might be most economically stored in ways ("soft constraints") which utilise gross topological similarities to features present at the retina, especially if we are allowing modelling by analogue computers. By "economically" I mean relative to fairly simple energy and interference constraints on cellular architecture such as those displayed by the optic nerve, which would seem in turn to require the observed neuroarchitecture of the primary visual areas of the cortex. The conclusion to draw here is that the conscious aspects of shape perception may simply *be* relatively coarse aspects of the spatial distribution of neural architecture, in line with Smolensky's proposition (17) but not his proposition (6). The homunculi of the somatosensory system suggest a similar story. Sometimes, in other words, "look to the neurons" is not such bad advice as Smolensky makes out.

A final point is that the interaction between neural, subsymbolic, and symbolic levels defeats any simple reductive thesis of everything to the subsymbolic level. Public communication at the symbolic level ensures its autonomy, for example. I emphasise also the ubiquity of the perceptual. The partial truth in empiricism pertains to the extent to which our concepts are permeated with the sensory: We cannot escape our prelinguistic evolutionary past. Hence, if the neural plays a greater role in the analysis of the sensory, it cannot be neglected in a proper account of other levels, and the completeness clause of Smolensky's proposition (8c) is thrown into question.

Connections among connections

R. J. Nelson

Department of Philosophy, Case Western Reserve University, Cleveland, Ohio 44106

I agree with most of what Smolensky says about the aims of a connectionist approach; nothing I have to say is meant to be critical of connectionist research in cognitive science. What I understand of it is impressive indeed. However, nothing Smolensky says convinces me that what he calls the "traditional cognitive model" can't in principle supply the theoretical power a connectionist model can, and more.

A lot depends on how we understand "traditional model." For Smolensky it is based on a symbolic paradigm: Processors manipulate discrete symbols and follow algorithms; they operate sequentially, that is, are von Neumann machines, and they mimic conscious rule interpretation. This characterization of the traditional model assumes that cognitive processes are modeled by *programs* operating on data, which is of course the common practice in artificial intelligence circles.

Von Neumann machine programs determine sequences of fetch-execute cycles using a single port memory. Programs are generally written in high level languages and translated to or executed from, machine language level, using data structures such as lists, trees, semantic nets, and the like. Cognitive scientists are the first to maintain that the mind is not very much like the computer model structurally and that computer processing is not strongly equivalent to the cognitive, but perhaps stronger than mere simulation. According to connectionists, another shortcoming is that the model doesn't seem to be of much use in explaining highly parallel cognitive activity or associative mnemonic processes.

The connectionist's is a pretty inadequate notion of contemporary processors, or of sequential processors themselves on the subprogram level. Multiprocessors, which are also connectionist systems, geometrically speaking, can perform tasks in parallel using simultaneously all the feedback you can imagine and more. Beyond that, there are multiaccess memory mainframes that have multi-ported memories (an example is the Cray X-MP-2) and fast wired-in means of comparison which, when combined with multi-ported memories, obviate associative memory architectures. These are not von Neumann machines and have not been extensively used, so far as I know, in AI. In none of the literature with which I am familiar have they been used for simulating connectionist schemes in cognitive science.

On the subprogram or logic level, a garden variety sequential computer is a discrete state system, i.e. a finite automation. Finite automata operate on decoded symbols (of an assembly system, for example) and hence its symbols are more finely grained than those presupposed in the "traditional model." Finite automata are parallel (of course, as Smolensky says, what counts as parallel and serial depends on the level of description). For example, a network realizing the $m+n$ relative recursions for the next state (f) and output (g) respectively of an FA

$$y_i(0) = k \qquad k = 0, 1$$

$$y_i(t+1) = f_i(x_1(t), \ldots, x_n(t), y_1(t), \ldots, y_m(t)) \qquad i = 1, \ldots, m$$

$$z_j(t) = g_j(x_1(t), \ldots, x_n(t), y_1(t), \ldots, y_m(t)) \qquad j = 1, \ldots, n$$

is about as parallel and interactive as you can get; every state element y is connected to and influences every other.

Restricting remarks to machine language programming, in the processors we know about (either sequential or parallel) there are two levels of algorithm guiding a process, the algorithms written in the program and the algorithm embodied in the circuit logic. For instance, a program contains the instruction ADD, and the logic level network, when obeying a command, follows a built-in algorithm represented by a set of functionals of the type displayed above. These relations can be written as production rules, as can the machine language program. So they are *both algorithmic in the same sense*, while they are manifestly different in architecture.

Thus the logic network level is subconceptual, subsymbolic, parallel, using finely grained representations and operates intuitively in the sense that in following a program imposed at the conceptual level it executes another at the logic network level, but not by anything remotely similar to "conscious rule interpretation." In addition, like connectionist models, logic networks (if considered for modeling cognition) are at a level intermediate between the symbolic conceptual level and the hardware or neural level. This idea of embodied algorithm is important in the application of Church's Thesis to cognitive science (Nelson 1987a).

Whether the mind is anything like this is of course relatively unknown. It is quite clear that net recursions, although connectionist for sure, are not very much like the connectionist nets of the cognitive scientist. Nevertheless, assuming computationalism is on the right track, the picture I have drawn is very similar to the connectionist's: Cognitive activity goes forward on a symbolic level modeled in the traditional way, and is associated in some way to a fine-grained subsymbolic, parallel, etc. process modeled by connectionist nets. I can say a lot more than that. A first-year computer engineering student could design a logic net realizing a given connectionist scheme using compositions of finite automata, i.e., ordinary discrete state methods, with the exception that the excitatory and inhibitory connectors among units of the network would bear associated integral rather than real values. I am not certain what limitation this would impose. But the automaton version has advantages (one of them is that there is no mystery in interlevel connectivity), and is far from being a "simulation" in the sense that a sequential program would be. This possibility shows both that the connectionist type of model is *theoretically* dispensable and is replaceable by a computationalist model that is by no means a simulation on a von Neuman machine. It's done by realizing the connectionist's heuristic model at a soft, logic level. I have little idea whether this is desirable, but it could be done.

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

But more than this, I follow Pylyshyn (1984) in suspecting that connectionist models are, qua explanations, of what he calls the "functional architecture" type. A necessary condition for cognitive modeling is that it appeals to the information-bearing content of representations. Put more formally, representations should be semantically interpretable essentially in the sense of model theory (or better, in the sense of some theory of reference that unfortunately does not exist yet); otherwise they are not fit operands of cognitive activity but just causal links in the functional architecture level. Furthermore, if Pylyshyn is right, connectionism cannot explain intentional attitudes.

I am aware that this view is subject to serious dispute. It is quite possible that representations or "subsymbols" are semantically interpretable – content bearing. However, most of the uses of "representation" are quite ambiguous (representation of information is not the same thing as symbolic content-bearing information; and neither are the same as semantic net representations of "meanings"). As this is so, I am not certain that I would understand any claim that Smolensky's subsymbols are content bearing in the relevant sense without first being instructed in connectionist terminology.

It can be argued, Nelson (1987b) against Pylyshyn (1984), that finite automata logic circuit nets do manipulate content bearing symbols, and moreover can in principle account for propositional attitudes, at least at the level of sensory expectations and perceptual belief (Nelson 1982). I accordingly advance the thesis that connectionist methods, though extremely valuable as heuristic tools, do not capture the distinctive qualities of cognitions and that logic nets – which are parallel, subsymbolic, operating below a conscious control level, and so forth – do.

A great merit of connectionist research beyond the interesting and fruitful experiments it has produced is its serving to put the program paradigm finally in proper perspective. Pattern recognition, for instance, is far more appropriately approached on the connectionist (either Smolensky's or a digital logic network level) than on the programming level. I suspect the same is true of language acquisition and elsewhere. This is hardly news; but the work of connectionist cognitive scientists helps make it sink in.

Subsymbols aren't much good outside of a symbol-processing architecture

Alan Prince[a] and Steven Pinker[b]

[a]Department of Psychology, Brandeis University, Waltham, Mass. 02254 and [b]Department of Brain and Cognitive Sciences, MIT, Cambridge, Mass. 02139

1. On the issues dividing connectionism and symbol systems. Smolensky's analysis relies on a series of spurious theoretical conflations:

1.1. Symbolic = conceptual, connectionist = subconceptual. For Smolensky, connectionist theories represent a radical departure because they invoke a subsymbolic or featural level of analysis, which contrasts with the consciously accessible concepts, easily labeled by words, that symbol-processing theories are committed to. But in fact symbolic theories have no a priori commitment to the "conceptual" level. Phonological distinctive features are a perfect example, and they were brought into generative linguistics for precisely the reasons that connectionists now embrace subsymbols: to define appropriate dimensions of similarity and generalization (Jakobson, et al. 1951; Halle 1962). Similarly, subsymbolic features are routinely utilized in syntax, morphology, and semantics.

This is presumably why Smolensky excludes formal grammar from "the symbolic paradigm." This is untenable. The "subsymbols" of linguistics are handled by rules and principles of an unmistakably symbol-processing type. Frequently discussed in the abstract as prototypical of the symbolic paradigm, linguistic theories are implemented as parsers in a variety of symbol-processing architectures. More concretely, connectionists' own models of morphology and syntax (Rumelhart & McClelland 1986; McClelland & Kawamoto 1986) bear scant resemblance to those of formal linguistics. In any case, "subconceptual" featural analyses can also be found in theories of nonlinguistic cognition (e.g., reasoning: Tversky & Kahneman 1983; vision: Marr 1982; Ullman 1984).

1.2. Parallel = connectionism; serial = symbol systems. An algorithm is serial to the extent that it requires an order of steps in its execution. The symbolic paradigm has no a priori commitment to strict seriality. Among linguistic theories, for example, some are highly serialized, others rely entirely on sets of conditions applying simultaneously. Less obviously, connectionism itself supports serial processing, when units are wired into feedforward layers, or, as in Smolensky's model, when "macrodecisions" or stabilizations of parts of networks occur in specific orders.

1.3. Context-sensitive = connectionism; context-free = symbol systems. Smolensky's example of the context-sensitive/context-free distinction relies on a fairy tale about the acquisition of the English past tense, which we discuss below. We note that contrary to Smolensky's assertion, grammatical rules do not have consequences "singly"; they take their meanings in the context of other rules, a fact at the heart of linguistic explanation. For example, the rules he cites – "past (*go*) = *went*" and "past (*x*) = *x*+*ed*" – have an intrinsic formal relation, due to the generality of the regular rule: since [go] is a possible instance of [x], either rule is capable of applying to *go*. This relationship is resolved by a principle that adjudicates between the general and special cases (see Pinker 1984). Adding a new rule, then, can radically change the ecology of a grammar. This is "context-dependence" of exactly the right sort, and it dispels the mystery that Smolensky sees in the fact that adults' *went* supplants children's *goed*.

1.4. Connectionist explanations are "exact"; symbolic explanations are "approximate." Smolensky implies that stochastic search at the microscopic level of a network provides an "exact" account of a cognitive process, whereas the structure of harmony maxima, while a convenient summary of the network's global behavior, describes the process only approximately. This conflates two distinctions: macroscopic versus microscopic levels of analysis, and exact versus approximate descriptions. Two descriptions could each be exact at different levels; it seems odd to say, for example, that an account of brain function in terms of neurons is only "approximate," with the "exact" account lying at the level of atoms. Different levels of analysis are motivated – both of them true "exactly" if true at all – whenever systematic events at the macro-level are not exhaustively predictable or motivated by principles of interaction at the micro-level.

This is exactly the case in many connectionist models: Far from being self-organizing, they are often wired by hand, and their parameters are tuned and tweaked, so that they behave properly (that is, they assume the global harmony maxima the theorist desires). This is not motivated by any connectionist principles; at the level at which the manipulations are effected, units and connections are indistinguishable and could be wired together any way one pleased. The question "Why is the network wired that way?" is answered by the macro-theory – "because phonological processes apply to adjacent segments"; "because the verb determines the role assigned to its object," and so forth. These answers are not "approximate." A successful symbolic theory may dictate the representations, operations, and architecture so robustly that any lower-level analysis that diverges from it – fails to "implement" it – will be a flawed approximation.

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

2. The adequacy of connectionist architectures. The argument that begins in a rejection of symbols for commonsense concepts and jumps to a defense of connectionism has a hole at its center: Without symbol-processing machinery, subsymbols don't do much good. Connectionist models that are restricted to associations among subsymbols are demonstrably inadequate. Consider these problems (For details, see Pinker & Prince 1988):

2.1. Distinguishing structural relations and similarity relations. In symbol systems, a featural decomposition is just *one* of the records associated with an entity: The feature vector can be selectively ignored by some processes, and the entity is represented by its own symbol, giving it an existence independent of the vector. In contrast, for a prototypical network using "distributed" representations, the entity is nothing but its features. For example, in the Rumelhart & McClelland (1986) model of past tense acquisition, a word is represented as a pattern of phonological features. This leads to an immediate problem: representing linear order. Since the vector must represent both the word's phonetic features and how they are concatenated, each feature must encode both kinds of information simultaneously. Thus such "Wickelfeatures" encode the presence of three adjacent phonological features; for example, "*unvoiced-unvoiced-voiced*" is one of the Wickelfeatures activated for *stay*.

Smolensky cites the Wickelfeature as a clear case of a "subsymbol." But note the bait-and-switch: Subsymbols were originally introduced as entities that are more abstract or fine-grained than symbols corresponding to commonsense concepts. But now they consist of features of an entity conflated with features of its context, combined into a single unit. This is necessary because the semantics of the features must do the work ordinarily carried out by symbol-processing architecture, in this case, preserving concatenative structure. The problem is that this move has disastrous empirical effects. Some words (e.g. *albal, albalbal* in the Australian language Oykangand) can't be represented uniquely because they contain several instances of a Wickelfeature, and the model can't count Wickelfeatures; it can only turn them on. It is difficult to explain psychological similarity: Wickelfeaturally, *slit* and *silt* have as much in common as *bird* and *clam*. The model can learn bizarre, nonexistent morphological rules (e.g., reverse the order of the phonemes of the stem) as easily as common rules (e.g., do nothing to the stem; add *d* to the stem). Some rules (e.g., reduplicate the last syllable) can't be learned at all.

This illustrates a dilemma inherent in Smolensky's program: Connectionists need to invoke subsymbolic features to get empirical successes, but their impoverished associationist mechanisms force them to use not the features demanded by the nature of the desired generalization, as revealed by macrotheory, but features that simultaneously carry burdens usually assigned to the symbol-processing architecture, such as preserving order. The subsymbols must do several jobs at once, none successfully.

2.2. Keeping individuals from blending. With feature-only representations, representing two things simultaneously is problematic. If early visual input is an activation pattern over feature maps, then without a serial attentional mechanism you can't tell the difference between a green circle near a red square and a red circle near a green square. In the Rumelhart & McClelland model, various subregularities and the regular rule all compete to activate Wickelfeatures for the past form: When *sing* is input, you could get features for *sing, sang, sung, singed, sanged,* etc., all superimposed. The model isn't putting out any single word at all, just an unordered collection of features, many of them contradictory. Thus it couldn't avoid blending what should be distinct competing outputs into bizarre hybrids such as *toureder* as the past of *tour,* or *membled* for *mail*.

It is easy to conceive of hypothetical languages in which speakers compose words by probabilistically superimposing bits of material into any one of a family of related combinations, depending on the frequencies of competing generalizations they have been exposed to. It is significant that human languages don't work that way. Connectionist networks, which superimpose the features of distinct individuals onto a single vector, leave this a mystery.

2.3. Distinguishing types and tokens. If objects are represented only in terms of their subsymbolic features, two objects with the same features get the same representation, and thus anything associated to one gets associated to the other. For the Rumelhart & McClelland model this raises many problems: *ring* and *wring,* for example, ought to go to *rang* and *wrung,* but the model can't enforce this difference because the past form is directly associated with the phonological representation of the stem. In traditional theories, past forms are associated with symbols representing the word itself, eliminating the problem. (Incidentally, adding semantic features won't help here.) Again, connectionist representations face conflicting demands, in this case, fostering generalization and keeping individuals distinct.

2.4. Selectively ignoring similarity. Similarities among individuals captured in their feature overlap must sometimes be shelved. For the past tense, phonological similarity plays a role in predicting forms within the system of irregular verb roots (cf. *sting/stung, cling/clung, stick/stuck*), but when verbs are derived from nouns (which clearly cannot be marked as having "irregular past tense"), phonological similarity goes out the window: You get *He high-sticked Lafleur,* not *high-stuck*. What is needed is some all-or-none mechanism that accesses featural information in some cases and ignores it in others. Nonlinguistic concepts can impose the same requirements.

2.5. Knowledge above and beyond trained associations. If the only available machinery is a set of units and connections, then the obvious way to learn is to connect more strongly those units that frequently co-occur in the input. Virtually all connectionist models learn according to this associationist doctrine. Smolensky reproduces Rumelhart & McClelland's version concerning the onset of overregularization in children: that it must be because the environment changes from a mixture of verbs in which irregulars predominate to a mixture in which regulars predominate. However, this is false: The ratio does not change. According to traditional symbolic explanations, the change occurs because the child memorizes past forms in the first stage and coins a rule capable of generating them in the second. All the data are consistent with this explanation. Thus we must reject Smolensky's argument that overregularization shows that cognition is nonmonotonic, radically context-sensitive, etc.

Attributing any of these phenomena to a separate rule-processor that is fed explicit culturally transmitted conventions is an illegitimate escape hatch. Not even the most fanatical yuppies will find a school that will instruct their child in the principles necessary to prevent him from saying *high-stuck;* nor do they have to. In fact, since the intelligence of connectionist models relies on specific input histories, whereas that of symbolic models relies on unconscious principles wildly unlike anything formulated in language curricula, explicit instruction will make people behave more like connectionist networks.

In sum: Smolensky conflates logically distinct contrasts in a way that stacks the deck in favor of connectionism. His defense of connectionism is the claim that good cognitive models will decompose commonsense concepts into more fine-grained, subtle, or abstract features or "subsymbols." But this claim is orthogonal to the connectionism debate. Prototypical symbolic theories, such as grammars, have always incorporated "subsymbols"; prototypical connectionist models, because they choose an associationist architecture, require their subsymbols to accomplish a set of mutually contradictory tasks and hence suffer as cognitive theories.

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

A two-dimensional array of models of cognitive function

Gardner C. Quarton

Mental Health Research Institute, University of Michigan, Ann Arbor, Mich. 48109

Imagine a large two-dimensional array of models (simulations) of cognitive processes. A family of closely related models is stored in a single column. At level one there is a simple model that takes as input some set of messages that represent some subset of plausible inputs to a human being engaged in a cognitive function and produces as output a set of messages that could, plausibly, represent the resulting output of the cognitive process. If this model is treated as a black box, we do not care, at level one, what computations go on inside as long as the output is determined in part by the input, and the "external behavior" of the black box bears some resemblance to human function.

At level two, within the same column, and, therefore, within the same family of models, we have an elaboration of the model at level one. Some of the details of events within the level-one black box are simulated as new black boxes nested within the level-one black box. This model specifies to a limited degree how the behavior of the model at level one is implemented. To make the model work at level two, we shall have to develop mechanisms for computing the output of the level-two black boxes from their input, but we do not treat this implementation as part of the model that mimics the human cognition. Note that we can implement the external behavior of a black box in two ways. One, which can be called "simulation relevant," requires the specification of a new layer of black boxes organized in a fashion that is part of the simulation. In other words, if we were checking the veridicality of the model, we would expect the external behavior of both the level-one and the external behavior of the level-two black boxes to mimic the cognitive process we are simulating. The second mode of implementation, which we shall call "simulation irrelevant," makes the external behavior of the black box be what we wish it to be, but it can do this in ways that we know are not likely to be similar to the way in which such behavior is implemented in the human function being modeled. In many models of cognitive function using von Neumann computers, the actual computations within the lowest level of black boxes may be of the simulation irrelevant sort. Of course, in some sense this implementation is not irrelevant because it makes the model work. However, it is irrelevant in the sense that we make no claim that this implementation mimics the comparable implementation in what we are modeling.

Now assume that the boxes in this column are filled in down to level one hundred. Each model except the level-one model is a further elaboration of the model above it, created by specifying nested black boxes, organized in a specified pattern, behaving in a specified way, and playing an implementation-relevant role. The black boxes in the innermost nest of level one hundred still, of course, are implemented in an implementation irrelevant fashion. Let us assume, however, that the level one hundred model is a simulation of a complete nervous system, and that the innermost black boxes represent molecules in membranes of neurons, and other entities at that level of biological detail. We have described just one column of the array. The other columns represent different simulating strategies. For instance, it may be that the first twenty columns represent modeling strategies based on the traditional symbolic paradigm. They are all different, but they share this feature. The next twenty are all connectionist. Still more columns may represent simulation strategies we have not thought of yet. It turns out, however, that many of the columns representing the symbolic paradigm behave plausibly at the upper levels, but as we move down the column it becomes increasingly difficult to pretend that the implementations resemble those in the human being. The human being after all does not run LISP programs. As a result, the cells of the columns below the uppermost levels are blank. This will also be true for the columns representing connectionist strategies, but it is possible that the models are plausible a number of levels below those of the symbolic paradigm.

Smolensky argues that "most of the foundational issues surrounding the connectionist approach turn, in one way or another, on the level of analysis adopted." He ends up suggesting that there are three levels of analysis, the symbolic paradigm at the top, the subsymbolic paradigm in the middle, and the neural level at the bottom. He defends his choice of three levels rather than two because he does not wish to suggest that connectionist modeling operates at the neural level.

I believe his three-level approach confounds the two ways of comparing models which are represented in my imaginary array by the two dimensions. Models differ in the strategy realized by their choice of implementing computation. They also differ in the degree of detail achieved by the simulation-relevant implementation and the degree to which this simulation-relevant implementation is supposed to resemble the function of the human nervous system.

Many neuroscientists would agree that connectionist models like those developed by Smolensky do involve more levels of simulation-relevant detail than do those in the symbolic paradigm. They may also agree that connectionist models seem to simulate not only the messages involved in the cognitive function but also the message vehicles, that is, the connections and the connection strengths that change dynamically over time. However, most of them would say that we are still a very long way from a veridical simulation of a human nervous system, and it is not clear whether this strategy is the beginning of a path to such a simulation or a blind alley.

We need detailed review papers that compare different modeling strategies (columns) at all the achieved levels of implementation (rows). We need such reviews for many different types of cognitive function (a third dimension?). Smolensky deserves a great deal of credit for realizing that a programmatic description of his strategy – independent of his actual models, but using them as illustrations – would permit an intensive review of many of the problems he faces. The other commentaries included here should help identify those issues needing more examination. Taxonomies of parallel processing computer algorithms and a mapping of these on computer architectures share some features with taxonomies of connectionist models. The members of these categories seem to adopt idealizing simplifications that make them too simple to be useful simulations of nervous system distributed information processing, parallelism, and concurrent computation. A much more detailed exploration of parallel processing (concurrent) algorithms, parallel processing computers, and connectionist models may be needed before neurophysiologists can develop the necessary new hypotheses.

Sanity surrounded by madness

Georges Rey

Department of Philosophy, University of Maryland, College Park, Md. 20742

Smolensky's account of connectionism is a mixture of positive and negative proposals. The positive ones (e.g., 8a and 8b) are generalizations of interesting results regarding specific cognitive processes; the negative ones (e.g., 8c) involve the rejection of certain claims of "symbolic" approaches.[1] Smolensky is careful (in claims (1a–e)) to admit the limitations of present connectionist results and to avoid dismissing the symbolic approaches out of hand. However, he also wants to avoid the "genuine defeat" of regarding connectionist models as "mere implementations" of symbolic ones. I want to locate here just

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

where – between eliminating symbolic approaches for connectionist ones, and reducing the one to the other – Smolensky's own position lies.

Symbolic approaches are attractive for a wide variety of reasons, chief among them their capacity to deal with the following phenomena:

(1) The structure of attitudes. There is a difference between thinking that someone loves everyone and that everyone is loved by someone or other; symbolic approaches capture this kind of difference by relating the agent to systematically different symbolic structures.

(2) The fine-grainedness of attitudes. There is a difference between thinking that Mark is fat, Sam is fat, that man is fat, and the funniest American writer is fat, even where Mark = Sam = that man = the funniest American writer. There is even a difference between thinking that a square lies in a circle and that a square lies in a locus of coplanar points equidistant from a given point. Symbolic approaches permit distinguishing these attitudes by distinguishing syntactically between different, even cointentional symbolic structures to which an agent can be related.

(3) The causal efficacy of attitudes. Ordinarily, someone thinking that someone loves everyone disposes the thinker also to think that everyone is loved by someone, but not vice versa; and ordinarily thinking that Mark is fat, and if fat then bald, can lead someone to think that Mark (but again not Sam, unless one thinks that Mark = Sam) is bald. Almost as ordinarily, people are biased toward positive instances in confirming hypotheses, ignore background frequencies in assessing probabilities, and are prone to falling into gambler fallacies. All these different patterns of thought cause people to behave in systematically different ways. If one supposes that the parts of structures needed in (1) and (2) are causally efficacious, symbolic approaches can capture both these rational and irrational patterns of thought.

(4) The multiple roles of attitudes. People often wish for the very thing that they believe does not presently obtain, for example, a drink of water, or that Sam (but not Mark) might come to dinner. Symbolic approaches capture this phenomenon by permitting different roles and access relations to the same symbolic structures.

Against all these reasons for symbolic approaches, there are the well-known problems that Smolensky cites regarding how "brittle" and "impractical" they are: their failure to capture the extraordinary swiftness of perception and thought, their failure to perform "gracefully" in degraded circumstances. Connectionist models do appear in these respects to be better. But, of course, their advantages in these respects will amount to little if they enjoy them at the expense of (1)–(4) above.

To some extent, Smolensky anticipates this issue. What advantages there are to symbolic models can be captured by regarding them as special cases of connectionist models: Symbolic structures are, for example, to be identified with "patterns of activation" in connectionist systems. These special cases, however, are "crude" ones: Connectionism ought in the end to replace approximate symbolic approaches, just as quantum mechanics ought in the end to replace classical physics.

Now, it is not at all clear to me how these patterns of activation will in fact be able to do all the work demanded by (1)–(4). Are the patterns structurally decomposable (e.g., into operators, quantifiers, connectives, terms) in the ways required by (1)? Can they be distinguished finely enough to capture the distinctions demanded by (2)? Are they and their parts available for the multitude of different relations and interactions required by (3) and (4)? It is possible that Smolensky has positive answers to these questions; or perhaps he has other ways of capturing these phenomena, or an argument that they are spurious. But he needs to present a great deal more discussion to make any of these possibilities – to say nothing of the bold claims of (1f–k) – even remotely plausible.

Suppose, however, that patterns of activation *can* be shown to play the role of symbolic structures. Why think that the latter structures are only crude approximations, that nonconscious processing is not tractable at the symbolic level, but "only" at the subsymbolic one? Smolensky's pessimism in this regard is no doubt based in part on the aforementioned problems of symbolic models in AI. But, notoriously, AI has been largely concerned with emulating human *behavior*. Someone might suggest we look instead for laws of a system's competencies. Why shouldn't we expect there to be symbolic laws capturing the competencies underlying, for example, (1) and (3)?

Smolensky worries about this issue as well. He acknowledges the competence/performance distinction, but reverses the usual understanding of it: Where the symbolic approach presumes that the laws will characterize competencies, performance being explained as the result of interactions, Smolensky expects the laws to lie with performance, competencies being explained as special cases. But this reversal alone can't be a problem, since a special case may still be a perfectly exact one. Where this difference in perspective makes a theoretical difference is in the way the speciality arises: The symbolic approach presumes competence laws will concern the *internal* states of the system, whereas Smolensky claims that competence laws will emerge only out of specific *environmental* conditions. The internal system by itself has no sharp conceptual order: From a conceptual point of view it is a hodgepodge of associations governed by nonsymbolic "thermodynamic" laws. Competencies are "harmony maxima" arising out of a general network of chaos: "If in the midst of life we are in death, so in sanity are we surrounded by madness," observed Wittgenstein (1956) in the midst of remarks on mathematics.[2]

This is an ingenious and to my mind improbable claim, added to the general connectionist approach. To make it plausible, Smolensky needs to show not only that connectionism can accommodate (1)–(4), but also that it will do so in essentially the same way that his system learned Ohm's Law, without internal symbolic laws emerging. I don't see how the example generalizes, however. People's ability to handle both valid and invalid inferences of the sort noted in (3) seems to be quite general and nongraduated: Once you see the forms you can apply them to an indefinite variety of cases; they do not seem to be stimulus-driven in the way that Smolensky's view requires them to be. But neither do I see that symbolic approaches are really in the end tied to one view of competence over the other, nor that connectionism ought to be so tied. Connectionist networks might still be interesting even if the more classical picture of competence and performance survived: Performance might often be the result of a network, for which a symbolic system is a fall-back.

In any case, why think that being an implementation of a symbolic system would be a "defeat" for the connectionist? Should it turn out that there are symbolic laws, but that symbolic structures can be encoded gracefully only as patterns of activation, this would be of considerable significance for both a connectionist and a symbolic approach. It might provide the requisite account of the speed with which symbolic structures are accessed, and of the role of stereotypes and "family resemblances" in much ordinary inference.[3] Each approach could then benefit from the strengths of the other. One needn't, after all, be "the only president you've got" to pique the interest of investigators.[4]

NOTES

1. I acquiesce here only for the sake of argument in Smolensky's presumption that connectionist networks are in some important way "nonsymbolic." The case has yet to be made that, in the ultimate (as yet unprovided) explanation of why the networks succeed, nodes in the network should not be taken to refer to various features of, for example, the stimulus. Smolensky's claim that they do not refer to features of which we are ordinarily *conscious* is quite beside the point; no symbolic story need make any such commitment. Perhaps the point for the time

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

being is this: The computational atoms in a connectionist network do not seem to be the syntactic atoms over which a compositional syntax and semantics are standardly defined. However, particularly in view of the possibility that "patterns of activation" could correspond to standard symbolic structures, even this weaker claim needs to be demonstrated.

2. Smolensky in this way adds a new perspective to an old debate, siding here not only with the later as against the early Wittgenstein (with regard to which see also Kripke 1980), but with Hume against Kant, Skinner against Chomsky, and, most recently, Burge (1986) and Barwise (1986) against Fodor (1986; 1987).

3. Which is not to say that the latter tricks exhaust ordinary inference. That many of our ordinary concepts, for example, exceed the stereotypes and resemblances that we may exploit in *accessing* them seems to be a further, rule-governed phenomenon that is not obviously amenable to a connectionist approach; for further discussion see Rey 1983; 1985.

4. I'm indebted to David Israel, Georg Schwarz, and Paul Smolensky himself for stimulating discussions of aspects of this topic.

Making the connections

Jay G. Rueckl

Department of Psychology, Harvard University, Cambridge, Mass. 02138

Among the many fundamental issues considered by Smolensky are the relationships between connectionist models and models at the symbolic and neural levels. Although I am generally in agreement with Smolensky, I would like to comment on each of these relationships.

The symbolic level. A crucial issue here is whether connectionist models should be seen as competing with symbolic models, or if instead connectionist models are merely implementations of symbolic models at a lower level of description. Smolensky explicitly rejects the implementational view (hypothesis 10, sect. 2.4.), and examines in some detail the points of incompatibility between the two frameworks (sections 2 and 5–9). Smolensky's arguments are compelling. Nonetheless, one might suppose that even though these frameworks are presently incompatible, they might eventually be made compatible through a process of coevolution. That is, developments at one level might bring about changes in the formulation of models at the other level, so that in the long run connectionist and symbol level models might be seen as isomorphic. For example, some of the attractive emergent properties of connectionist systems, such as content-addressable memory and incremental learning, might be taken as primitives in symbolic level models that are assumed to be implemented on connectionist architectures (Hinton, McClelland & Rumelhart 1986; Oden 1987).

This would be a happy outcome, but there are reasons to doubt that it will occur, and I would like to supplement Smolensky's arguments by pointing out one problem that seems particularly difficult to overcome. The problem (touched on by Smolensky) concerns the discrete character of computation at the symbolic level. For example, in the typical symbolic model instances get assigned to categories in an all-or-none fashion. Similarly, production rules and other sorts of computational processes are executed when discrete conditions are met, and have discrete results. Recent work has shown that symbolic models can be "fuzzified" to some extent. Category membership can be made a matter of degree, and logical operators that retain fuzzy information can be defined (Oden 1977). Similarly, production systems can take into account the degree to which the conditions of a production are satisfied, and the strength of the action taken can depend on the degree to which the rule's conditions were met (Anderson 1983).

One might imagine a way of identifying fuzzy symbols with distributed patterns of activity, thus bridging the gap between the symbolic and subsymbolic levels of description. For example, one might equate the degree to which the conditions of a production are satisfied with the degree to which a certain pattern of activity is present. The problem with this approach is that knowing the degree to which a pattern is present is not sufficient for predicting the behavior of the system. One must also know which parts of the pattern are or aren't present. If, in a connectionist model, a given pattern of activity in module A causes a related pattern of activity in module B, there is no guarantee that all module A patterns that overlap with the key pattern to the same degree will result in the same module B pattern. Under certain mapping functions the various B patterns could be wildly different. The point is that patterns of activity at the subsymbolic level are representations with causally efficacious internal structures. Symbol level descriptions, fuzzy or not, lose that internal structure, and thus seem destined to fail to distinguish between certain causally distinct states.

The neural level. Smolensky compares neural and connectionist architectures along a variety of dimensions (Table 1 in target article), and the lesson he draws from this comparison is that the subsymbolic and neural levels are conceptually distinct. Thus, Smolensky argues, the subsymbolic level has a sort of autonomous existence. The implication is that although it would be nice to make connections between the subsymbolic and neural levels, there is plenty of work to be done at the subsymbolic level alone, and this work should not be subject to arguments concerning neural implausibility.

Although I agree with Smolensky's arguments in principle, I think it is a mistake to emphasize the autonomy of the subsymbolic level while at the same time downplaying the potential for making deep contacts between theories at the subsymbolic and neural levels. A variety of considerations suggest that the attempt to make connections between these levels should be given high priority. First of all, those of us who have bought into the computational theory of mind are committed to the assumption that a bridge between the computational and neural levels of description exists, and we must thus expect that sooner or later an understanding of the connection between these levels will be a part of psychological theory. Second, discovering which computational algorithms are used by humans and other animals is hard work, and it would be foolish to ignore any information that might inspire the development of new kinds of algorithms or help to choose between alternatives under consideration. Work at the neural level has produced a wealth of such information, and those of us working at the computational level would do well to take it into account. (Indeed, see Kosslyn 1987, for an excellent example of a computational theory motivated in part by neuropsychological and neuroanatomical findings.) Third, developing connections between neural and computational models is likely to benefit neuroscientists as well as psychologists. As Smolensky points out, one reason that findings at the neural level have had relatively little impact on cognitive modeling is that, although we have a great deal of data about the brain, "these data are generally of the wrong kind for cognitive modeling" (sect. 4, para. 13). Although this is surely true to some extent, part of the problem is that computational models have typically ignored questions of neural instantiation, and thus have failed to generate empirical questions for neuroscientists to explore. By constructing theories that explicitly suggest how algorithms might be instantiated, theorists at the computational level might generate empirical predictions for neuroscientists to test. The results of these tests would in turn influence work at the computational level. This interplay between the two levels could only be of benefit to us all.

Conclusion. Smolensky's analysis of the relationship between connectionist models and models at the symbolic and neural levels seems on target. Smolensky suggests that the relationship between connectionist and symbolic models is similar to that between quantum and classical mechanics. Symbolic models and classical mechanics offer approximate descriptions of their respective domains, but fail in ways that can be understood within the connectionist and quantum frameworks, respec-

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

tively. Furthermore, the differences between the two accounts of each domain are fundamental, and in neither case can one theory be reduced to the other. I concur with Smolensky's analysis, and have offered one more reason to believe that symbolic models cannot be reduced to subsymbolic models.

I also concur with Smolensky's analysis of the relationship between the subsymbolic and neural levels. However, in this regard we have different visions of how cognitive science should proceed. Although he agrees that in the long run subsymbolic and neural models should be connected in a principled way, Smolensky stresses the autonomy of the subsymbolic level, and does not push for an increase in the interplay between research at the neural and computational levels. While I agree that work at the subsymbolic level will be fruitful regardless of the degree of contact with the neural level, I also suggest that attempts to make contact between the levels will be well worth the effort.

Structure and controlling subsymbolic processing

Walter Schneider

Learning R&D Center, University of Pittsburgh, Pittsburgh, Pa. 15260

The proper evolution of connectionism should relate multiple levels of description of cognition to constraints and mechanisms that affect each level of description. At present most connectionist processing involves associations between a set of input patterns and a set of output patterns. It has made important contributions: showing the interrelationships among patterns (e.g., Rumelhart, Smolensky, McClelland & Hinton 1986); developing more powerful learning rules (e.g., Rumelhart, Hinton & Williams 1986); and exploring the use of weights in representational and memory systems (e.g., Hinton & Plaut 1987); see also Schneider & Detweiler 1987. In general it has done so with an extremely limited space of connectionist architectures and processes. There have been exceptions (e.g., Touretzky 1986; Schneider & Detweiler 1987; and Smolensky 1987). However, most connectionist models are similar to NET talk (Sejnowski & Rosenberg 1986) in that there is an input layer, an output layer, and zero to two intermediate layers. The units are simple, quasilinear components summing the inputs with a possible threshold or logistic output function. This is a very simple architecture compared to the brain. Although these simple multilayered systems are useful model paradigms, an exploration of a richer set of architectures is called for.

Most connectionist modeling does not make contact with the structural or dynamic constraints of physiology. Smolensky remarks that neurophysiology provides the "wrong kind" of information for connectionism – providing structure rather than dynamic behavior. We do know a fair amount about the structure of the brain (e.g., see Van Essen 1985) and this information can be used to identify connection patterns for complex computation (e.g., Ballard 1986; Schneider & Mumme 1987). We also have information about the dynamics of the system (e.g., that minimal neural activation times are in the range of 5 to 50 milliseconds; and attention can modify a signal by a factor of 3, but requires 60 milliseconds to occur [Moran & Desimone 1986]). Connectionism needs to examine a richer class of connective structures and modulatory processes. This richer class raises questions such as: What is the effect of heterarchical connectivity (as in Van Essen 1985)? How do multispeed learning rates (Mishkin, et al. 1984; Hinton & Plaut 1987; Schneider & Detweiler 1987) influence working memory? And what computational advantage is there in using an attentional control structure (Schneider & Mumme 1987)?

A richer set of architectures may show that symbol processing is more than an emergent property of connectionist vector processing. Smolensky (sect. 2.4., [10]) faults symbol processing when it suggests that connectionist processing is a low-level implementation of symbol processing. Smolensky claims that symbol processing is an emergent property of connectionist processing. This claim seems premature. Some properties, such as categorization of symbol-like entities (e.g., J. A. Anderson & Mozer 1981) are clearly emergent. Some properties, such as variable binding, require a whole control architecture of processing components (e.g., gating cells, binding cells) to maintain, compare, and copy activation patterns (e.g., Schneider & Detweiler 1987; Smolensky 1987; Touretzky 1986). These are *not* emergent properties; rather, they are hand crafted to perform population-based processing activities that produce symbolic-like processing. It is likely that as connectionist modeling expands from the limited associative mapping paradigm, a plethora of connectionist modules will be needed to accomplish extensive symbolic processing. Both connectionist and symbolic processing can make important contributions to an understanding of these behaviors. Rather than claiming that one level is the emergent or implementation version of the other, it would be better to identify the weaknesses and strengths of each and examine hybrid architectures that can better cover the space of human behavior.

Smolensky suggests that symbol (S-knowledge) and pattern (P-knowledge) exist in one connectionist medium. This is possible, but it may be that they are quite different processes, implemented very differently in the architecture. Symbolic learning often occurs in a single trial (see J. R. Anderson 1983). In contrast, connectionist learning typically occurs in the time scale of thousands and sometimes millions of trials (see simulations in Rumelhart, McClelland & the PDP Research Group 1986). Human behavior exhibits qualitatively different types of behavior (see Shiffrin & Schneider 1977) when these two types of knowledge are being used. Single-trial learning typically results in slow, serial, effortful processing, whereas extended consistent practice produces relatively fast, parallel, low-effort processing. If the single-trial learning is done via specialized bind-cell processing (e.g., Touretzky 1986) one is no longer in the same medium. The bind cells can be built out of connectionist hardware or Turing machines. They operate on a meta-level above the connectionist vector processing hardware. In a model of human attentional processing (Schneider & Mumme, forthcoming), connectionist populations perform categorization and association operations. This allows the execution of P-knowledge; however, it takes hundreds of trials to develop reliable associative patterns. On top of an architecture of connectionist modules, a control mechanism is *not* emergent from the associative input/output processing, but rather from a new processing element to moderate the interactions that occur when multiple messages need to be multiplexed serially to limit crosstalk. The control level can itself be implemented in connectionist hardware. Since control processing operates at the meta-level, its activation modulates populations at the lower level. This provides a symbolic-like control structure modulating the vector transmissions. This control process can acquire rules in a single trial by maintaining the condition-action pairs in vector modules. Input vectors can be compared; if there is a match, the action vector is transmitted. As training progresses (over hundreds of trials), the input vector becomes associated to the output vector, allowing direct input to output association without the use of the control processing.

Connectionism is a major advance in the modeling of cognition and has already had a significant impact on psychology (see Schneider & Detweiler 1987). However, it must become a member of a team of concepts and tools for the study of cognition, rather than trying to produce a paradigm shift supplanting its predecessors. A wide range of architectures should be explored in trying to cover a space of human behaviors while using available physiological, behavioral, and computational constraints. Neurophysiologists tell a story that if you can think of five ways that the brain can do something, it does it in all five,

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

plus five you haven't thought of yet. In the study of cognition we need to control our desire to have one answer, or one view, and work with multiple views.

How fully should connectionism be activated? Two sources of excitation and one of inhibition

Roger N. Shepard

Department of Psychology, Stanford University, Stanford, Calif. 94305-2130

Smolensky develops a persuasive case that connectionism provides a significant level of description between the level of conceptual processes accessible to introspection and verbal communication, and the level of neural processes probed by physiologists. Advocates of the symbol manipulation approach to cognition (as well as advocates of the ecological approach to perception) may facetiously suggest that connectionism is thereby "filling a much needed gap" in the explanatory hierarchy. But, for me, connectionism has two exciting features that have been lacking in the discrete, symbolic, propositional theories that have dominated cognitive science.

The two sources of excitation. First, connectionism offers a dense, massively parallel processing medium that in addition to facilitating ties to the neuronal substrate appears more suited to subserving such analog processes as apparent motion and imagined transformation (Cooper 1976; Shepard & Cooper 1982; Shepard & Metzler 1971).

Second, connectionism promises to furnish, for what has been a largely ad hoc approach to cognitive modeling akin to that of engineering, a more deeply principled ground akin to that of physics. Instead of simulating human capabilities by larger and larger patchworks of heterogeneous, domain-specific heuristics, connectionism seeks a uniform framework within which diverse performances and competences arise from a small set of general principles. I am thus heartened in my own quest for a kind of Newtonian mechanics of mind that may be governed by "universal" psychological laws (Shepard 1984, 1987).

However, the general principles so far put forward by connectionists concern only the first two of the following three processes needed to achieve adaptive behavior.

Three processes of adaptation. *Inference:* On the shortest time scale, upon encountering a particular situation, there is the process of adapting the internal representation and overt response to the requirements of that situation – even though no situation is ever completely revealed in the available sensory input. In a connectionist system, perceptual completion, interpretation, categorization, prediction, and inference are achieved by the passage, through state space, of the vector specifying the momentary levels of activation of all elements in the processing network to a stationary vector (or "eigenstate"), in accordance with what Smolensky terms the "activation evolution equation" (sect. 2.3., para. 5) (formalizing, perhaps, relaxation methods for the satisfaction of "soft constraints"). The set of situations giving rise to the same stationary vector correspond, in "psychological space," to what I have recently termed a "consequential region" (Shepard 1987).

Learning: On an intermediate time scale, over a series of encounters with situations from some ensemble, adaptation to the ensemble is achieved through principles governing the slower passage, through weight space, of a vector specifying the strengths of all connections between the elements in the processing network, in accordance with what Smolensky calls the "connection evolution equation," formalizing, perhaps, "backwards error propagation" (sect. 2.3., 7; footnote 7). Through such "tuning," an initially chosen vector of weights comes to determine increasingly refined trajectories and final states for the vector of activations and, hence, increasingly effective inference.

Selection of initial structure: On the longest time scale, the topology and initial weights of a connectionist network are the result of some evolutionary process that generates systems with different connectivities and weights and eliminates those systems that fail to learn to draw inferences appropriate to a particular ensemble of situations. For living systems, this process is that of mutation and natural selection. For artificial systems, it has been a more haphazard and idiosyncratic one of guess and test. In either case, the imposed connectivity and initial weights determine what inferences are learnable and what sequences of situations are sufficient for such learning.

A source of inhibition. Smolensky rightly observes (sect. 1.1, para. 3) that "much of the allure of the connectionist approach" is that through tuning of their own weights, connectionist networks "*program themselves.*" But nontrivial self-programming can take place only if some knowledge about the world in which the system is to learn is already built in. Any system that is without structure has no basis for generalization to new situations (Shepard 1964; 1981; 1987).

Smolensky also rightly emphasizes (sect. 7.1) that the purpose of the "subsymbolic" system must be to achieve a "veridical representation of ... environmental states, with respect to ... given goal conditions" (para. 1). However, in focusing on the achievement of such representations through the two processes of inference and learning, he (like other connectionists) seems to slight what I regard as the most challenging problem of cognitive science, namely, the problem of the source and internal form and operation of innate constraints.

I distinguish internal representations of particular external objects from internalizations of general constraints that have governed all such objects and their transformations throughout evolutionary history. Particular foods, predators, or places of safety or danger, having varied from one locale or epoch to another, could not be internalized as innately fixed knowledge and are largely learned. However, the invariable constraints, such as that relative light/warmth alternates with relative dark/cold in a 24-hour cycle, or that space is locally Euclidean and three-dimensional, can be shown to have led to the internalizations of a circadian clock and an intuitive grasp of kinematic geometry that are probably innate (Shepard 1984).

Smolensky is quite explicit about the difficulty of characterizing representations at the subconceptual level, which does not preserve the semantics of the consciously accessible conceptual level (sect. 5). Commentators have sometimes voiced the objection that even if a connectionist system manifests intelligent behavior, it provides no understanding of the mind because its workings remain as inscrutable as those of the mind itself. The force of this objection is mitigated if the *principles* of learning and inference that govern the internal representations can be explicitly stated – even if the form of the internal representations themselves cannot. However, because significant structure must be built into the system before effective learning and inference can take place, we face two alternatives: Either we must formulate how the required structure is to be implemented at the inscrutable subconceptual level; or we must formalize explicit principles for the evolution of connectionist systems analogous to the principles of learning in individual systems.

ACKNOWLEDGMENT

Preparation of this commentary was supported by National Science Foundation Grant Number BNS 85-11685, while Shepard was the Fowler Hamilton Visiting Research Fellow at Christ Church College, University of Oxford.

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

From connectionism to eliminativism

Stephen P. Stich

Department of Philosophy, University of California, San Diego, La Jolla, Calif. 92093

Smolensky's portrait of connectionism is a welcome and exciting one. The burden of my commentary will be that if the project he describes can be carried off, the consequences may be much more revolutionary than he suggests. For if it turns out that Smolensky-style connectionist models can indeed be constructed for a broad range of psychological phenomena of both the "intuitive" and the "consciously conceptualized" sort, then, it seems to me, a pair of very radical conclusions will plausibly follow. The first is that folk psychology – the cluster of commonsense psychological concepts and principles that we use in everyday life to predict and explain each other's behavior – is in serious trouble. The second is that much psychological theorizing that cleaves to what Smolensky calls the "symbolic paradigm" is in serious trouble as well. In both cases, the trouble I envision is the same: The theories are false and the things they posit don't exist. Since space is limited, I'll limit my remarks to theories in the symbolic paradigm, and leave folk psychology for another occasion.

A central thesis in Smolensky's rendition of connectionism is that a "complete, formal account of cognition" does not "lie at the conceptual level" but at the "*sub*conceptual level" (sect. 2.4, para. 5). Earlier, in making much the same point, he tells us that "complete, formal and *precise* descriptions of the intuitive processor are generally tractable not at the conceptual level, but only at the subconceptual level" ((8)c, sect. 2.3, para. 7, emphasis added). But what exactly does Smolensky have in mind when he claims that a complete, formal, *precise* account of cognition is to be found only at the subconceptual level? As I read him, what Smolensky is claiming is that the real, exceptionless, counterfactual supporting generalizations or laws of cognition are only to be found at this level. At the conceptual level, by contrast, such generalizations as we have will be at best rough and ready approximations that may be more or less accurate within a limited set of boundary conditions, and generally not very accurate at all when we go outside those boundary conditions. If this thesis turns out to be correct, then the cognitive states and processes posited by connectionist models will be the ones describable by genuine laws of nature, but there will be no laws describing the doings of the semantically interpreted mental symbols posited by theories at the symbolic level. If we want accurate predictions of the phenomena, they will have to be sought at the subsymbolic level. As Smolensky would be the first to agree, the thesis he sketches is at this point only a hopeful guess. To defend it requires that connectionists actually build models for a broad range of phenomena, and demonstrate that they do indeed yield more accurate predictions than competing models at the conceptual level. But let us assume that the thesis will ultimately be established, and consider the consequences for theories and posits at the conceptual level.

To start us off, an analogy may prove helpful. For Lavoisier, in the last quarter of the 18th century, heat was caused by caloric, an "exquisitely elastic fluid" "permeating all nature, which penetrates bodies to a greater or lesser degree in proportion to their temperature" (quoted in Gillispie 1960, p. 240 & p. 239). When Sadi Carnot formulated the second law of thermodynamics in 1822, he "still handled caloric as flowing from a real reservoir of heat down a continuous gradient" (Gillispie 1960, p. 241). For many years the theory of heat that posited caloric was embedded in an evolving, progressive, sophisticated research program that generated both explanations of observed phenomena and increasingly accurate predictions. Ultimately, however, that theory was rejected and replaced by the kinetic theory. Though the detailed history of this transition is a complicated story, a crucial factor was that the new theory sustained more accurate predictions and better explanations over a broader range of phenomena. Moreover, since the kinetic theory posits no "exquisitely elastic fluid," and recognizes no laws governing its flow, those who were prepared to grant that the kinetic theory is better concluded that caloric theory is false, and that the fluid it posits does not exist.

Consider now the analogies that will obtain between this case and the case of conceptual level psychological theories if Smolensky's thesis turns out to be right. Like caloric theory, the conceptual paradigm has sustained an evolving, progressive, sophisticated research tradition. But if Smolensky is right, we will find that the generalizations of conceptual level theories (like those of caloric theory) are only approximations and apply only in limited domains, while the generalizations of subconceptual level theories (like those of kinetic theory or its successors) are "complete" and "precise." Against the background of this analogy, it is tempting to conclude that if Smolensky's thesis is right, then conceptual level theories are false, and the entities they posit do not exist.

There is reason to suppose that Smolensky himself would not resist the first half of this conclusion. For at one point he tells us that "the relationship between subsymbolic and symbolic models is . . . like that between quantum and classical mechanics" (sect. 5, para. 11). But, of course, if quantum mechanics is right, then classical mechanics is wrong. Whatever its virtues, and they are many, classical mechanics is a false theory.

The second half of the conclusion I'm trying to coax from my analogy is the more distinctively eliminativist half. (For some background on "eliminativism" see P. M. Churchland 1894, pp. 43–49; P. S. Churchland 1986, pp. 395–99; Stich 1983, Chapter 11.) What it claims is that the entities posited by conceptual level theories are like caloric in one very crucial respect; they do not exist. From his one brief mention of "naive . . . eliminative reductionism" (sect. 10, para. 2). I'd guess that Smolensky would be more reluctant to endorse this half of my conclusion. Nor would such reluctance be patently unjustified. For it is certainly not the case that whenever one theory supplants another we must conclude that the entities posited by the old theory do not exist. Often a more appropriate conclusion is that the rejected theory was wrong, perhaps seriously wrong, about some of the properties of the entities in its domain, or about the laws governing those entities, and that the newer theory gives us a more accurate account of those very same entities. Thus, for example, pre-Copernican astronomy was very wrong about the nature of the planets and the laws governing their movement. But it would be something of a joke to suggest that Copernicus and Galileo showed that the planets Ptolemy spoke of do not exist. So to defend the eliminativist half of my conclusion, I must argue that the connectionist revolution, as Smolensky envisions it, bears a greater similarity to the rejection of the caloric theory than to the rejection of geocentrism.

In arguing the point, it would be useful if there were, in the philosophy of science literature, some generally accepted account of when theory change sustains an eliminativist conclusion and when it does not. Unfortunately, however, there is no such account. So the best we can do is to look at the posits of the old theory (the ones that are at risk of elimination) and ask whether there is anything in the new theory that they might be identified with. If the posits of the new theory strike us as deeply and fundamentally different from those of the old theory, in the way that molecular motion seems deeply and fundamentally different from "exquisitely elastic" caloric fluid, then the eliminativist conclusion will be in order. Though, since there is no easy measure of how "deeply and fundamentally different" a pair of posits are, our conclusion is bound to be a judgment call. That said, let me offer a few observations which, I think, support a proeliminativist judgment.

Smolensky notes, quite correctly in my view, that in the

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

dominant approach to cognitive modeling (the approach that he calls the "symbolic paradigm") symbols play a fundamental role. He goes on to note that these symbols have a pair of fundamental characteristics: They refer to external objects and they are "operated upon by 'symbol manipulation'" (cf. sect. 1.3., para. 3). Smolensky does not elaborate on the idea that symbols are operated on by symbol manipulation, but I take it that part of what he means is that, in the models in question, symbol tokens are assumed to have a reasonably discrete, autonomous existence; they are the sorts of things that can be added to, removed from or moved around in strings, lists, trees and other sorts of structures, and this sort of movement is governed by purely formal principles. Moreover, in the symbolic paradigm, these sorts of symbol manipulations are typically taken to be the processes subserving various cognitive phenomena. Thus, for example, when a subject who had previously believed that the hippie touched the debutante comes to think that the hippie did not touch the debutante, symbolic models will capture the fact by adding a negation operator to the discrete, specifiable symbol structure that had subserved the previous belief. Similarly, when a person acquires a new concept, say the concept of an echidna, symbolic models will capture the fact by adding to memory one or more symbol structures containing a new, discrete, independently manipulable symbol that refers to a certain class of external objects, namely echidnas.

In connectionist models, by contrast, there are no discrete, independently manipulable symbols that refer to external objects. Nor are there discrete, independently manipulable clusters of elements (or "subsymbols") which may be viewed as doing the work of symbols. When a network that had previously said yes in response to "Did the hippie touch the debutante?" is retrained to say no, it will generally not be the case that there is some stable, identifiable cluster of elements which represent the proposition that the hippie touched the debutante, both before and after the retraining. And when a network that was previously unable to give sensible answers to questions about echidnas is trained or reprogrammed to give such answers, there typically will not be any identifiable cluster of elements which have taken on the role of referring to echidnas. Instead, what happens in both of these cases is that there is a widespread readjustment of weights throughout the network. As Smolensky notes, the representation of information in connectionist models (particularly in parallel distributed processing style models) is widely distributed, with each unit participating in the representation of many different aspects of the total information represented in the system. This radical disparity between strategies of representation in symbolic and PDP models makes a smooth reduction – or indeed *any* reduction – of symbols to elements (or to patterns of activity) extremely implausible. Rather, I submit, the relation between mental symbols and connectionist elements (or patterns of activity) is akin to the relation between caloric and molecular motion. If this is right, then in those domains where connectionist models prove to be empirically superior to symbolic alternatives, the inference to draw is that mental symbols do not exist.

From data to dynamics: The use of multiple levels of analysis

Gregory O. Stone

Department of Psychology, Arizona State University, Tempe, Ariz. 85281

While focusing on the substantive differences between connectionism and traditional cognitive science, Smolensky's analysis illustrates a fundamental epistemological difference. In the traditional approach, the symbolic level is the "correct" level of analysis. Other levels, such as hardware implementation, are effectively considered irrelevant. In contrast, Smolensky argues that "successful lower-level theories generally serve not to replace higher-level ones, but to enrich them, to explain their successes and failures, to fill in where the higher-level theories are inadequate, and to unify disparate higher-level accounts." Thus, connectionism may portend a revival, in cognitive science, of theoretical pluralism – the philosophy that no single perspective can fully account for observed phenomena (James 1967).

Smolensky presents three levels of analysis (neural, subconceptual, and conceptual) as a priori theoretical constructs. I will argue, however, that the choice of levels derives from a strategy of maximizing the explanatory power of the pluralistic framework in which they are embedded. In other words, levels of analysis are primarily pragmatic constructs.

What are the advantages of a pluralistic methodology? One common objection to connectionist models is that their complexity hinders an understanding of what they are doing and why. This conceptual opacity is, to some extent, a price paid for their flexibility and generality of application, allowing models built from a few basic mechanisms to account for a broad range of disparate phenomena. On the other hand, mechanisms explicitly tailored for specific operating characteristics tend to be limited in their generality. This often leads to a profusion of unconnected, but eminently testable and transparently interpretable, special-purpose models. A methodology which uses and interrelates both levels of analysis can exploit the strengths and overcome the weaknesses of each when considered in isolation.

A concrete example will help to clarify this point. Reeves and Sperling (1986) asked subjects to report, in order, the first four items (digits) from a rapidly presented visual sequence. But subjects first had to shift their attention from another part of the visual field to the digit stream. The attention shift altered the perceived order of items in the sequence, producing an inverted-U shaped recall function. In the first phase of their analysis, they found that a scalar precedence or order score for each item in each condition provided a very powerful account of the data. However, this analysis invoked a large number of parameters and offered no conceptual insight into why the observed precedences were obtained. Their second level of analysis produced a close fit to these precedence scores by treating them as the result of the temporal integration of an item's input strength. A slow-opening attention gate reduced the input strength of early items, which lead to the inverted-U shape of precedence scores across position. This level of analysis provided a conceptual framework with greater parsimony; however, it was domain-specific and provided no link to temporal order in short-term memory in the absence of an attention shift.

Grossberg and Stone (1986) extended the analysis to the subconceptual level by mapping the Reeves and Sperling model into the short-term memory dynamics of adaptive resonance theory. The analysis began with an abstraction from the extremely complex activation dynamics to an emergent and more tractable functional form relating the relative precedence strengths. This emergent functional form is necessary for stable, long-term encoding. When this functional form was applied to the Reeves and Sperling model, several unexpected principles of short-term memory dynamics and attentional gain control were revealed. Furthermore, the experimentally derived order scores were accounted for using a mechanism which plays a critical role in adaptive resonance theory treatments of other short-term memory phenomena, as well as treatments of categorization, unitization, and contextual facilitation (see articles reprinted in Grossberg 1987a; 1987b). The key point in this example is that it would have been difficult – if not impossible – to have achieved the same degree of insight by mapping the data directly into the class of possible short-term memory dynamics.

Each level of analysis in the preceding example served an important role in the overall methodology.

The descriptive level of analysis encapsulates the raw data in a

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

more tractable, but relatively atheoretical form. Computational overhead is reduced in the search for mechanisms and functional forms with optimal explanatory power, and a descriptive model can reveal fundamental structure underlying the data.

The functional level of analysis expresses structure in the data in terms of high-order conceptual constructs. At this level, one develops functional characterizations of processing, such as temporal integration of input strength. This is a potentially broad class of analysis, of which the symbolic paradigm is a special case.

The dynamic or subconceptual level of analysis provides mechanistic details underlying the broad constructs of the functional level. What appeared to be simple, domain-specific, processes are now seen as the subtle interaction of many more general mechanisms.

Perhaps the most important component of the methodology is the development of overarching design principles that interrelate the levels of analysis and govern their theoretical development. Because functional constructs arise from subtle, nonlinear interactions between dynamic mechanisms, alteration of a single mechanism can affect the performance of the whole system. As a result, previous predictive capability can be lost in some attempt to introduce new predictive capabilities. Design principles identify the critical features of a mechanism responsible for a desired operating characteristic. If a mechanism must be redesigned, previously identified principles remain to guide the process; one need not begin again from scratch. Design principles provide a conceptual bridge between dynamic mechanisms and functional constructs, and thus help elucidate what a dynamic system model is doing and why.

Work remains to be done in developing a powerful, pluralist framework for cognition and behavior. In particular, much of the methodology currently used by both the symbolic and the connectionist paradigms will need to be replaced or reworked. Smolensky's insightful investigation of fundamental assumptions is an important contribution to the development of this framework. Unless the development of a pluralist methodology continues, connectionism will fail to achieve its great explanatory potential.

ACKNOWLEDGMENT

Thanks to Lex Akers, Ron Epperlein, Scott Hill, Don Homa, Peter Killeen, Guy Van Orden, and Sharon Walker for helpful comments on early drafts.

On the proper treatment of thermostats

David S. Touretzky

Computer Science Department, Carnegie Mellon University, Pittsburgh, Pa. 15213

"Eliminative" connectionists (Pinker & Prince 1988) are the radicals of the connectionist movement. They make the boldest claims with the least evidence. The contribution of Smolensky's target article is that it eloquently states (and even numbers) the points of faith that define the eliminative stance. I cannot prove Smolensky's PTC (proper treatment of connectionism) wrong, but I believe its principal and most radical claim, that formal symbolic theories of intelligence will turn out to be inadequate for explaining human performance, is very badly in need of some supporting data. The problem is most noticeable with respect to language.

Hypothesis (16) of PTC assigns responsibility for language to an intuitive processor, while (8a–c) reject the notion that this processor might be implemented in the brain as a formal sequential rule interpreter. Many nonconnectionists share this view. But in the collective restatement of (8a–c) as (8), PTC makes its far stronger, radical claim: that it is impossible *in principle* to give an accurate account of intuitive phenomena at the symbolic level. Such an account can be achieved at a lower, nonsymbolic level, we are told. This is where eliminativists get themselves into trouble.

PTC is not a competence theory, it is a performance theory. On the other hand, PTC is supposed to be more abstract than the neural level; it is not obligated to explain every hesitation and every muscle twitch. What sorts of performance phenomena might PTC account for that symbolic theories cannot? In any physical system there are bound to be insignificant jitters that can only be explained by going to a lower level of description. In order for PTC to be confirmed, its supporters must be able to demonstrate *significant* linguistic effects that do not admit symbolic-level explanations. This introduces two themes for debate: Which performance effects are significant, and which of those are not covered by symbolic-level theories?

Consider a thermostat with setpoint T_0 whose behavior is determined by the following rule:

IF $T < T_0$ THEN turn-on(furnace)
ELSE turn-off(furnace)

It makes no difference that the thermostat has no symbols and no rule interpreter inside it; the above rule is a description of the thermostat's behavior that any cognitive scientist would feel comfortable with. It is a formal rule because it generates precise predictions and can be implemented in various ways. It is a symbolic-level rule because it is expressed as relationships among the terms that actually define the domain: ambient temperature, setpoint, and furnace activity. It makes no reference to mechanisms or processes whose behavior is unrelated to the domain description.

Now suppose that as the switch inside the thermostat closes, it bounces a few times, causing a brief oscillation in the output signal. Furthermore, imagine that the furnace which the thermostat controls emits heat the instant it is told to do so. Our hypothetical thermostat-furnace combination therefore produces a few milliseconds of temperature oscillation whenever T drops below T_0, followed by a steady temperature increase until T_0 is again exceeded. But the formal account of the thermostat's behavior says nothing about bouncing switch contacts, because that is an implementation detail that has nothing to do with the domain. Therefore it cannot explain the oscillation.

The temperature oscillation problem may be dealt with in several ways. (*i*) Classify extremely short-duration phenomena as irrelevant on teleological grounds (the thermostat's purpose is broad-timescale temperature control, not microregulation) or psychophysical ones (people who choose the thermostat's setpoint are not sensitive to the oscillations). In either case the oscillations are excluded from the data. (*ii*) Discount the oscillations as performance error: They may be detectable, but need not be explained by a competence theory of temperature control. (*iii*) Declare the oscillations relevant and significant, and redo the formal account. This leads to a more complex behavioral rule for the thermostat, based on the difference between the current time and the time the ambient temperature most recently dropped below the setpoint. Because it is a symbolic level theory it still says nothing about bimetallic switches, and thus does not explain the cause of the oscillations; it merely reproduces them. (*iv*) Declare that in reality there is no formal symbolic level, therefore it is impossible to give a fully adequate account of thermostatic temperature control at this level of description. Instead, we should model thermostats as systems of contacts and springs and develop equations to describe their switching dynamics.

Which of these strategies is the correct one? Each has been applied to some aspect of language: (*i*) excludes nonlinguistic phenomena such as breathing patterns while speaking; (*ii*) relieves us of the need to account for ungrammatical utterances in a competence theory; (*iii*) argues for the recognition of what are obviously implementation-dependent limitations, such as

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

people's inability to handle deeply center-embedded sentences. In other words, (*iii*) shifts the emphasis from competence to performance, which, as Smolensky notes, is in part what distinguishes cognitive theories from linguistic ones. The eliminative heresy of PTC is (*iv*); it is also espoused by Rumelhart and McClelland (1986a; 1986b).

Even a diehard eliminativist would agree that the significant aspects of thermostat behavior can be described perfectly well at the symbolic level. The fundamental question regarding PTC and language is whether the effects attributable to the underlying dynamical system are as trivial and incidental as switch bouncing, or are instead profound, influencing the very structure of our linguistic facility. If the latter is the case, those who follow approach (*iii*) will be forced to produce bizarre, contorted rules in order to give a symbolic account of performance phenomena that PTC can explain quite naturally. Due to the limited scope of current connectionist models, there is not yet any convincing evidence that this will in fact happen.

Even if the eliminative hypothesis is correct, why should we rely on relaxation as the dominant metaphor for subsymbolic computation? Simple dynamical systems are attractive because they are mathematically tractable, but if connectionists really expect to unravel language, the jewel of cognition, they had best give up the idea of doing it with either statistical mechanics or heteroassociators. This rules out virtually all distributed connectionist models to date, for example, Rumelhart and McClelland (1986b); McClelland and Kawamoto (1986); Sejnowski and Rosenberg (1987); and Allen (1987). As any defender of the symbolic-level paradigm would argue, connectionist models with persistent internal state (Jordan 1986; McClelland, personal communication), modular structure (Derthick 1987a; 1987b; Touretzky & Geva 1987), and built-in mechanisms for complex operations such as variable binding (Touretzky & Hinton 1985; Dolan & Dyer 1987) stand a better chance of success.

Connectionists have been exploiting *tabula rasa* learning and simple physics analogies like the proverbial drunk searching for his keys under a lamppost: because that's where the light happens to be. There is also plenty of light under the formal symbolic lamppost favored by traditional cognitive scientists. Perhaps the keys are lying in the shadows.

ACKNOWLEDGMENT

This work was supported by the Office of Naval Research under contract number N00014-86-K-0678.

The essential opacity of modular systems: Why even connectionism cannot give complete formal accounts of cognition

Marten J. den Uyl

Vakgroep Psychonomie, University of Amsterdam, 1018 XA Amsterdam, Netherlands

There can be no doubt that Smolensky has done an excellent job at unravelling some of the conceptual knots that connect the symbolic and subsymbolic paradigms in cognitive science. I find myself in close agreement with much of what Smolensky has to say about the promises connectionism holds for deepening our understanding of the human mind. Yet there is one issue where I strongly disagree with the views espoused by Smolensky.

Let me begin my argument with the observation that cognitive systems that perform complex tasks tend to adopt a modular architecture; the more complex the tasks and the wider the range, the more inevitable the assumption of modularity appears to become. Modular information processing is a well-known concept in theories of computation and it has recently come to play an important role in cognitive theories due to the work of Fodor (1983; see also multiple book review, *BBS* 8(1) 1985). We might accordingly expect connectionist models also to adopt modular architectures. A large majority of the work in connectionist modelling is in fact concerned with connectionist modules, that is, with small, specialized, and relatively encapsulated parts of some hypothetical larger processing structure. Typically this encompassing structure is represented only by way of some input/output and bookkeeping routines that feed and control the connectionist module.

Modular architectures present some complications for connectionist theories and it seems that Smolensky systematically underrates the importance of these problems. I believe that because of this neglect a confusion of levels shows throughout his exposition. The confusion is not between levels of description, which Smolensky goes at great length to disentangle, but between *levels of aggregation*. The point is that at various places Smolensky appears to assume that the characteristics of a single connectionist module may be transplanted unmodified to an extensive processing structure consisting of many intricately interrelated modules. For example, in his Table 1 (sect. 4), Smolensky presents some relations between neural and subsymbolic architectures. It strikes me that many of the discrepancies Smolensky observes – e.g. "uniformly dense connections," "simple topology of distal projections between node pools" – have their source mainly in an inappropriate comparison between a single, structurally homogeneous, connectionist module and an extensive neuronal structure that quite likely supports an intricately modular processing structure.

If we take the human cognitive system to be a huge processing structure consisting of many interrelated connectionist modules, it follows that we may distinguish two different domains for connectionist theorizing: In one domain the primary concern is with the development of models for within-module processing, that is, with models that optimally perform the satisfaction of "soft constraints"; the second domain is primarily concerned with the development of theories of the modular structure of connectionist models, that is, with analyzing forms of interaction between connectionist modules. Obviously, this distinction can only be drawn very roughly at present, since there exist many interdependencies between the two sets of problems. (It may be noted that the distinction parallels in part, and only in part, the distinction between "distributed" and "local" connectionist theories.) It would seem that Smolensky's characterization of connectionist theorizing – e.g. the predominant role it assigns to continuous mathematics – holds nicely for theories of within-module processing, which may indeed be the domain of connectionism proper.

The situation is much less clear when we consider the theoretical domain of between-module interactions. It would seem that there are cases where the interactions between modules can be adequately described by the same kinds of differential equations as used in analyzing within-module processing. Probably, however, this will only work for "modeling simple aspects of cognition-like relative times for naming words in various contexts, or the relative probabilities of perceiving letters in various contexts" (Smolensky, sect. 2.3., para. 8).

Before we turn to more interesting cases of modular interactions, I will attempt a more specific interpretation of the general notion of a connectionist module. In the context of Smolensky's harmony theory (Smolensky 1984a; 1984b) it seems most natural to identify a module with a subset of units that cooperate simultaneously in achieving a "best fit set of inferences" (sect. 9.) – a highest harmony completion – over a part of a larger network. It is further implied that the process of "simulated annealing" is spatially bounded by a harmony module. It follows naturally that a module is closed or "opaque," that is, not passing activation to other modules, as long as it is in a state of "high computational temperature," when its internal activation pattern is highly erratic.

The discontinuities in activation passing that result naturally from modular structures are not in themselves insurmountable

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

obstacles for continuous analysis. However, a new level of complexity is introduced, where the interactions between modules take the form of passing not one-dimensional quantities of activation but discrete patterns of activity between modules. Continuous mathematical formalisms lose their appeal quite drastically when interactions between modules involve interactions between complex patterns of activity (cf. Smolensky, sect. 2.3, 8b).

Examples of cognitive tasks where these forms of intermodular interaction seem to be required in order to arrive at adequate connectionist models are ubiquitous. The most conspicuous case is "conscious rule interpretation." Touretzky's "BolzCon" model, which takes some important first steps toward implementing this human capability in a subsymbolic model (Touretzky 1986), is a highly modular system, using more than its share of patterned intermodular interactions. Another example is qualitative judgment, the striking capability of the human mind for fast, global, evaluative judgments. I have argued elsewhere (den Uyl 1986) that this capability can be best modeled in a connectionist system by assuming a specific modular architecture involving patterned interactions between stimulus and resonance patterns.

Why, then, should we reject the position that "the complete formal account of cognition . . . lies at the subconceptual level"? If my account of modularity in connectionist systems is basically correct, then the connectionist proper may still hope to devise "complete and formal" accounts of isolated connectionist modules. However, cognition is an attribute that may characterize systems as a whole, not single modules. The connectionist determined to study the performance of complete modular systems must accept that many of the most interesting behaviors cannot be adequately captured by mathematical formalisms.

Has the case been made against the ecumenical view of connectionism?

Robert Van Gulick

Department of Philosophy, Syracuse University, Syracuse, N.Y. 13210

Smolensky explicitly rejects any blandly ecumenical views about the relation between connectionism and traditional approaches to the symbolic modeling of cognition. In particular, he rejects the suggestion that connectionist systems fit within the standard framework as models of lower-level cognitive processes and as implementations of more conventional AI programs. Instead, he believes that connectionism offers an alternative symbolic model of higher-level cognitive processes, one that conflicts with the traditional approach. He asserts that the symbolic and subsymbolic paradigms for modeling cognition are strictly incompatible; they involve mutually inconsistent commitments.

Of course no one favors bland ecumenicism. We all enjoy a good intellectual fight, and a conflict of competing paradigms will probably produce more progress than would a harmonious accommodation that blurs the differences among diverse approaches. However, I find what Smolensky has to say about the incompatibility of the two paradigms less than convincing. Moreover, at points, especially in his concluding section, he seems to express decidedly ecumenical sentiments himself, despite his earlier claims.

The case for the incompatibility of the two paradigms is made most directly in Section 2.4., though even there it is qualified to some degree. Smolensky distinguishes between intuitive cognitive processes and those which involve conscious rule application. It is only with respect to the former that he offers an argument for strict incompatibility. He wishes to reject principle (10), which states that valid connectionist models are merely implementations of symbolic programs. His reason is that (10) contradicts (8c), the subconceptual level hypothesis, which he takes to be a key element of the subsymbolic paradigm. The two principles, as formulated, are indeed strictly inconsistent. However, the conflict concerns only the level at which exact, complete, and precise formal explanations of behavior are possible. Principle (8c) asserts that such descriptions of the intuitive processor will be possible only at the subconceptual level; (10) asserts that there are such descriptions at the conceptual level. All the work in generating a conflict between the two is being done by the demand for an *exact, complete,* and *precise* formal description of the intuitive processor.

But there are several problems with that demand. First, it is not at all clear just what is being demanded. How are "exact," "complete," and "precise" to be understood? Smolensky does not say. Second, and more important, in so far as it is clear what is meant by these terms, it is far from obvious that we should expect an exact, complete, and precise formal description to exist at any level of the intuitive processor. Rather, one might expect all such formal models to be at best approximately instantiated by the actual neural hardware. However, if one drops the demand for precise and complete formal description, the strict contradiction between the two paradigms disappears. There is no reason why the subsymbolic connectionist model, which is approximately instantiated by the neural structure at some level of description, might not be an implementation of a symbolic model, which is instantiated at a higher level of description. Indeed, Smolensky seems to accept this possibility with respect to those cognitive processes that involve conscious rule application. And he allows that even with respect to intuitive processors, conceptual level descriptions will be crudely approximated by the subsymbolic models he prefers.

I suspect that the important question is not, "At what level are complete precise formal descriptions possible?" but rather, "At what level will we find powerful insightful generalizations that help us understand the basis of cognition?" The questions are distinct, since interesting insightful generalizations about the formal nature of cognition need not involve formal models that are complete, exact, and precise. Smolensky uses the relation between the macroscopic description of a gas and its underlying microstructure to illustrate his view of the relation between symbolic and subsymbolic models. But the example might have a moral quite other than the one he intends. Though the regularities describing the emergent properties of the gas may be less precise and exact than those governing the mechanical interactions of its microscopic constituents, they may be the interesting or important ones for many explanatory purposes. Indeed as Putnam (1975) has argued, one should expect this to be the case when the macroproperties and macroregularities are relatively invariant across substantial variations in the underlying microstructure. Conversely the microstructure, like the structure of subsymbolic processes, is most important when variations in its properties greatly affect or constrain the nature of macrostructural (or conceptual level) regularities.

Thus it is probably best to embrace the nonreductionist and seemingly ecumenical viewpoint to which Smolensky turns in his concluding section. One should explore the nature of cognition at many levels of description, recognizing that in some cases the interesting regularities will be at the higher levels of description, but that in others they will be found at the subsymbolic level in structures and processes that have their own distinctive regularities, which can be no more than very incompletely understood in terms of the "shadows" or "images" they cast on the conceptual level of description.

One last caveat. Though there is value in emphasizing the diversity of competing approaches, one should not create differences where none really exist. For example, contrary to what Smolensky implies, the standard symbolic models of cognition make regular use of nonmonotonic reasoning (all you need is a commitment to a total-evidence condition) and processes which have a semantics other than that used to define the task domain.

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

It may be that some such features "come for free" in connectionist models, but whether free or otherwise they are certainly present in standard symbolic models.

The reality of the symbolic and subsymbolic systems

Andrew Woodfield and Adam Morton
Department of Philosophy, University of Bristol, Bristol BS8 1TB, Great Britain

Smolensky's picture of how the symbolic and subsymbolic levels are related runs together a thesis about two systems and a claim about the relative explanatory capacities of two paradigms. It would be simpler to theorize about the relations between levels for which one claims ontological reality, without at the same time trying to locate oneself in a paradigm.

Consider the following two-systems hypothesis: There are two different abstract types of system which the human brain can be taken to instantiate, symbolic and subsymbolic. Given a particular brain that instantiates both, how might the activities of those two system-tokens be related?

Case (1) *Division of labour.* Just as the agent may play the piano with one hand while stirring coffee with the other, so the cognizer may perform symbolic operations with one part of his brain while doing subsymbolic computation with another part.

Case (2) *Killing two birds with one stone.* Just as an agent can, in one arm movement, both signal a turn and wave to a friend, so the cognitive agent might, via one set of neural events, simultaneously manipulate symbols at the conceptual level and perform a subsymbolic operation. Two cognitive processes, functionally independent of one another, happen to be corealized in a very versatile physical substrate.

Case (3) The "by" relation. Just as an agent may enter into a contract *by* signing his name, so a cognizer may carry out a symbolic process *by* performing subsymbolic operations. Two levels of cognition are both mediated by the same neural events, but they are not functionally isolated. The symbolic process *emerges out of* the subsymbolic, or is "level-generated" by it (Goldman 1970). A connectionist machine can, under certain conditions, simulate a von Neumann machine. Smolensky suggests, analogously, that human beings, unlike digital computers, might be cognitively hard in virtue of being soft machines that have attained a high level of complexity. It is worth emphasizing that if symbolic thinking is indeed an emergent, it really exists, just as the act of entering into a contract exists. You cannot be emergentist *and* eliminativist about hard processing. The position we offer as a foil to Smolensky's is that over a significant range of cognitive tasks, the human brain functions as a symbol-manipulator *by* being a subsymbolic system.

A complication is introduced by the suggestion that human symbolic calculation might only *approximate* hardness. This gives rise to a new thought about reduction, and an analogy. Smolensky suggests that "symbolic" theories may be reducible to "subsymbolic" theories in roughly the same way that classical mechanics is reducible to quantum mechanics. This analogy with physics, while illuminating, captures neither the ontological commitment to the two levels nor the suggestion that the symbolic system in humans arose historically out of a new arrangement of preexisting capacities. Analogies for developmental emergence are rare in physics, but plentiful in biology. For example, ethologists have proposed that some social behaviour in animals (display, courtship rituals) evolved out of displacement activities produced by conflict between basic instincts.

Second, the fact that theories at the symbolic level are only approximately true of humans would not undermine the ontological commitments of the "two systems" hypothesis. The brain might really instantiate a symbolic system that was type-identified via an idealized specification. Human performance usually falls short of perfection. To succumb to "as if" locutions, or to say that there is basically only one system after all, is incompatible with the "emergence" thesis. Smolensky sometimes appears to be tempted in this direction. To claim, however, that the human brain instantiates *nothing but* a subsymbolic system prevents one from saying, as Smolensky wishes to do, that the theory of subsymbolic processing explains how the human brain *does* manage to instantiate a symbolic system.

Third, although the subsymbolic level may explain why certain symbolic level descriptions are true, the direction of explanation could equally well go the other way. From certain perspectives, explanations at the symbolic level could be more basic than explanations at the subsymbolic level. Consider again the analogy with action-theory. A child approximately succeeds in the act of eating jelly with a spoon, by moving his hand, his head, his mouth. Although the movement-description is lower-level than the act-description, the act-description is explanatorily more basic in a teleological sense, since we can explain why he performed that sequence of movements by reference to the act. Similarly, a piece of clumsy reasoning might be explained either in terms of the goals and norms that the thinker was trying to satisfy, or in terms of the subsymbolic processes that mediated his actual performance.

Why does Smolensky see his two "paradigms" as rivals? One reason, perhaps, is that he is over-reacting to an unfair, derogatory charge made by some High Church computationalists, that connectionist systems are "mere implementations" of systems whose "proper" level of description, qua cognitive, is symbolic. Such a charge has no sting. For one thing, there are probably many kinds of cognitive processes which do not use a language-like system of internal representation. More important, for processes that do require such representations, there is nothing "mere" about discovering that the processor in question emerges out of a connectionist system. To establish a convincing example of a case (3) relationship would be a great victory, and a vindication of the PDP (parallel distributed processing) approach. We learn *how* the symbolic processor does its job, but we also gain deeper insight into *what* is being done, from both a psychological and an evolutionary perspective.

Yet not even global success of this sort would prove that "the complete formal account of cognition lay at the subconceptual level," for this phrase presupposes *one* formal account that is complete. There might instead be two formal descriptions, each being a complete account of its own system, but neither giving a complete account of cognition as a whole.

Editorial Commentary

Some senses of "level" seem relatively clear and well defined: the hierarchy of compiled programming languages and the software versus hardware levels; the function/structure dichotomy; molar/molecular or macro/microlevels of description; performance/mechanism, behavior/neural substrate. All these seem to involve a viable higher/lower distinction. Perhaps conscious/unconscious processes can also be said to stand in some sort of superordinate/subordinate relation, although this begins to impinge on unsettled aspects of the mind/body problem. But what about "symbolic/*sub*symbolic"? Has an up/down relationship that is informatively called "levels" really been picked out here? On the face of it, symbolic and *non*symbolic appear to be the only two relevant options – at least if one is committed to an explicit formal definition of "symbolic" such as Fodor's [*BBS* 3(1)] or Pylyshyn's [*BBS* 3(1)] – but this is just parasitic on the software/hardware distinction, with no obvious intermediaries. To attempt to flesh it out by defining yet another up/down relation – "conceptual/*sub*conceptual" – seems either

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

to be parasitic in turn on the conscious/unconscious dichotomy or to declare the existence of an intermediate level by fiat. Similar remarks can be made about the subsymbolic/neural relation. Have different levels, or different senses of level, been conflated, or perhaps invented, in Smolensky's treatment of connectionism? For if we have no prior interpretative commitments, connectionism simply appears to be a simulated or implemented family of statistical algorithms for adjusting connection strengths in an interconnected causal network whose performance capacity and limits remain to be explored.

Author's Response

Putting together connectionism – again

Paul Smolensky

Department of Computer Science and Institute of Cognitive Science, University of Colorado, Boulder, Colo. 80309-0430

Table 1. *The format of this response. The commentaries discussed in each category (sometimes in footnotes) are listed in order of appearance*

1. Levels of analysis

The major issue discussed in the target article was the levels of analysis used in various approaches to cognitive science. Several of the commentaries misconstrued the PTC (proper treatment of connectionism) position on this issue, either explicitly or implicitly. The target article focused on what the PTC account *is;* I will now devote more attention to what it is *not,* providing a better framework in 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 on that of Pinker & Prince 1988).

The first possibility is the most straightforward: μ is an implementation of M. The notion of implementation is provided primarily by the von Neumann computer; throughout this discussion I will take "implementation" to mean exactly what it means in that context.[1] For my own 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 performing exactly 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 a complete, formal, and precise account of the system's computation.

If μ implements M, then this constitutes the strongest possible sense in which μ and M could both be 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); 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 there is no systematic relation between them. 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 2 presents the implementationalist and elim-

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

Table 2. *A spectrum of positions on connectionism's relation to the symbolic approach*

	Position				
	Eliminativist		Limitivist		
	Neural	Connectionist	(PTC)	Revisionist	Implementationalist
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)

inativist positions, along with a number of other relevant positions on the relation between connectionist and symbolic accounts. All positions in Table 2 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. As Table 2 indicates, and as I will shortly discuss, PTC adopts a view in some sense intermediate between the far-left eliminativist and far-right implementationalist views. The intermediacy of the PTC position allows us to understand an interesting phenomenon that occurred in the commentaries. Commentators 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 direct at PTC their favorite attacks on the opposite extreme. Thus we see why **Touretzky** – whose connectionist models (e.g., 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 (e.g., **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 more aspects of cognition than 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).

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 *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 nontrivial sense that I describe the micro- and macrotheories 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 occurring 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 microaccount μ and a macroaccount 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

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

principle (AE) through the proof (or less rigorous logical argument) that, given that the laws μ hold at the microlevel, it follows that the laws of M hold at the macrolevel. 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 correspond to a limit of relativistic quantum theory in which, loosely speaking, masses of bodies approach infinity and speeds approach 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: The number of connectionist units or the strength of connection approaches infinity (allowing "soft" properties to become "hard," and allowing memory capacity limits to be idealized away for "competence" theory), the relaxation or learning time approaches infinity (allowing, e.g., stochastic inference or learning to converge to theoretically known limits), and the overlap (inner product) of vectors stored in memory approaches zero (orthogonality: eliminating interference of items in memory).

Since the formal relationship between the micro- and macrolevels presumed here is one of convergence in the limit, the PTC position in Table 2 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 in C.

Having discussed the main points of Table 2, let us consider each of the positions outlined in the table, from the extreme left to the extreme right.

There are two kinds of eliminitivists: The furthest left position maintains that the science of cognition will require 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 to the 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 that it is taken into consideration that they are approximate and restricted; the range of cognitive phenomena modeled by the subconceptual accounts, and the exactness of those models, is much greater. In 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): *a)* describing consciously mediated (nonintuitive) processes – including many of the phenomena so important to philosophers, such as conscious reasoning; *b)* describing isolated aspects (i.e., not complete accounts) of performance; *c)* giving general (as opposed to detailed and formal) ways of understanding and thinking about cognitive processes; *d)* describing intuitive competences: abstractions away from performance (i.e., not precise), e.g., in language processing (contrary to **Rey,** the validity of competence theories is 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 a complete, formal, and precise account of cognition at the macrolevel. A favorite way to imagine the revisionist scenario playing out is to modify symbolic theory by relegating certain processes to connectionist networks: e.g. 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 symbol-like processing while the right half does soft, squishy, connectionist-like 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 (personal communication) and I call this arrangement by its French name, *cohabitation.* (**Woodfield & Morton** call this "division of labor.")

A subtler, but vaguer, revisionist view anticipates a 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 1988). 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 and 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

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

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 simply 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 (e.g., 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 does not 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, *cohabitation*, 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: "Which 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 1. 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

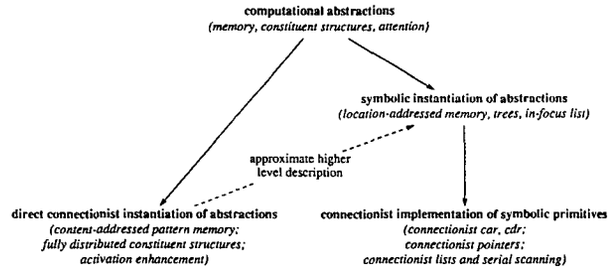

Figure 1. The methodological implications of limitavist (left branch) vs. implementationalist (right branch) views of connectionism.

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, whereas the left, PTC branch requires a reconstruction of the cognitive architecture in which the basic computational abstractions acquire new, nonequivalent, instantiations. This view of the relation between connectionism and cognitive architecture has much in common with that of Fodor and Pylyshyn (1988), but our assessments differ.

As the last two paragraphs show, there are important methodological implications that depend on 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 complete precise formal cognitive descriptions are to be found," for it is this question that determines which branch of Figure 1 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 1. 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

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

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, p. 569–73), 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 basic 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 go beyond 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 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 right 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 the EDITORIAL COMMENTARY and **Quarton**'s take my use of "symbolic" and "subsymbolic" to refer to levels. In fact, these terms refer to paradigms for cognitive modeling, *not* levels. The editor is right to question whether "subsymbolic" refers to a lower level than "symbolic": It does not. 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 (target article) illustrates, the symbolic/subsymbolic distinction is orthogonal to the distinction between the conceptual and subconceptual level. These 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, whereas the subconceptual level is comprised of fine-grained entities beneath the level of conscious concepts. For connectionist models the conceptual level consists of patterns of activity over many units and their interactions; 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 the EDITORIAL COMMENTARY 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 (target article), **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. Quarton 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 interpretation 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 preferred 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

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

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 there would in principle be 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's 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 a 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 eliminativist 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 (e.g., subjecting the thermostat to rapid temperature oscillations that prevent it from equilibrating).

As stated earlier, several implementationalist-leaning commentators mistook PTC for eliminativist; in addition to **Touretzky,** these include **Rey** ("connectionism ought in the end to replace . . .") and **Schneider** (". . . paradigm shift laying waste 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 2 other than the PTC position.

Dyer and **Touretzky** emphasize the importance of symbolic processes (e.g., 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, p. 1). Although 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 1, 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 marshal the power behind symbolic computation.

The commentators who seem to advocate revisionist positions include **Chandrasekaran et al.** ("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, preferring to view symbolic theory as an analog of caloric theory rather than of Newtonian mechanics. The moral he wants to draw from his analogy is that the microtheory (kinetic theory) eliminated the macrotheory from science. This 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 microtheories always eliminate macrotheories from science that I referred to in the conclusion of the target article as "naive . . . eliminative reductionism," and Stich is right 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 (e.g., 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 2: 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 accordingly. How can the causal powers of symbols be instantiated

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

analogously, without some system that recognizes relevant subsymbolic activity and acts upon that recognition? Or is that exactly what is being proposed?

Chandrasekaran et al. propose to characterize the levels issue in terms of Marr's (1982) 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. [See also Anderson: "Methodologies for Studying Human Knowledge" *BBS* 10(3) 1987.] 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 focused more exclusively on the lower level. Harmony theory, for example, can be accommodated 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 et al. are right to point out how understanding a model at the higher levels greatly promotes the 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 concerns relations between levels within a single model; it will not do, however, for the *between-model* relation that PTC posits 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 by 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 2 by arguing for what amounts to the Extremist Fallacy, which they state in their concluding paragraph. Their argument is that either connectionism denies that symbolic entities (e.g., 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 show briefly 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 and Levine offer no response, and their argument relies critically on refuting or ignoring this crucial point.

1.5. The neural level. The pursuit of increased coupling between the neural and subsymbolic modeling is advocated by **Lloyd, Mortensen, Rueckl,** and, to a lesser extent, **Bechtel.** My intent 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 close; but there are reasons to believe (see footnote 7 and preceding discussion in the target article) that in fact the opposite is now occurring. 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.

Den Uyl makes the important point that Table 1 (target article) 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 Uyl** 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 their structure and complexity. This conclusion may be correct, but it seems reasonable to adopt the working hypothesis that the mathematics describing current connectionist networks, if these are to be the modules of future systems, will have to contribute significantly 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. Whereas I accept the centrality of statistical inference to the paradigm, and its role in characterizing rationality, it seems too restrictive to exclude other computational processes

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

in dynamical systems (e.g., motor control) which have yet to be shown to fit in 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 turn out that 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 in 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 explanations of various cognitive regularities? Put differently, the "subconceptual level" hypothesized by the target article can be viewed as containing a number of sublevels, all lying between the conceptual and neural, all characterized by connectionist processing, lower 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 is successful: 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 (Dreyfus & Dreyfus, in press); discovering the form of an individual expert's knowledge and scientifically investigating the domain are almost the same activity. This is clearly not the case in the subsymbolic paradigm – unless we are prepared for a radically new definition of "scientifically investigating the domain." Belew goes on to point out that the connectionist approach places more weight on the dynamic properties of cognitive systems than on their static structural properties. A clear and simple example of this important point is the case of memory retrieval: In a traditional symbolic architecture, whether or not 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 focused too heavily on dynamical systems with simple static equilibria, and paid too little attention to dynamical systems with much more complex global behavior. Although this is undoubtedly true, it is changing, with Freeman's work on chaotic equilibria and Jordan's (1986) on periodic equilibria (where "equilibria" really means "attractors"). As for the flamboyantly eliminativist assertions ending his commentary, I think Freeman would be much harder pressed to live with the consequences of this neuromacho 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 in important ways from some "language of thought" accounts. This issue is discussed in the target article in Section 7.2, but Dreyfus & Dreyfus are right 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 refer to these levels. Nonetheless, his substantive comments seem by and large to 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 – however, the interpretation is in terms of subconceptual features such as "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 were looking in the target article for 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 (e.g., **Chandrasekaran et al., Lycan, Prince & Pinker, Rey, Van Gulick**).

There is a good reason why 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, broadly construed, which in turn can be used to implement a Turing machine, and so all of symbolic computation.

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

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

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 (e.g., **Chandrasekaran et al., Lindsay, Lycan, Prince & Pinker, 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.

That semantic levels and types of computational processes are independent is again indicated by the two-dimensional format of Table 2 (target article). 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 is 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 ground 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 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 did not willingly accept the constraints that connectionism imposes.

Although 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 (e.g., Anderson 1983) in terms of internalization of taught rules, followed by rule compilation and chunking, start with taught rules that must rely on consciously accessible concepts, and then manipulate these rules in ways that never go below the conscious level toward the conceptual. [See Anderson: "Methodologies for studying Human Knowledge" *BBS* 10(3) 1987.] Philosophical arguments from the structure of mental states, like those championed by Fodor and Pylyshyn (1988) 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. [See Chomsky: "Rules and Representations" *BBS* 3(1) 1980.]

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.

It seems that **Prince & Pinker** do not accept my right to define the grounds of my analysis to exclude linguistic-based models; they go on to accuse me of conflating a number of issues. I am perfectly aware that symbolic

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

computation can incorporate subconceptual features, parallel processing, and (as they might have added) soft constraints, nonmonotonic logic, and fuzzy sets. The irrelevance of all this for the present treatment of the target article has just been spelled out in the preceding paragraphs.

In reference to my remarks comparing inference with soft and hard constraints (Section 8), **Prince & Pinker** 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 a 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.

Prince & Pinker go on to enumerate 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 and McClelland (1986) model of past-tense acquisition. But it has been emphasized in the connectionist approach for some time (e.g., Hinton 1981), that it is in general 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 from consisting in part in a pattern encoding the phonological form and in part a pattern that serves as a unique identifier for that stem (e.g.. to distinguish homonyms). In fact, arguments such as those in the commentary of **Dreyfus & Dreyfus** imply that such microfeatures are to be expected among those invented by networks in their hidden units.

Next, **Prince & Pinker** 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.

Prince & Pinker are quite right 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 one, 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 on which Prince & Pinker base their entire critique.

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

Prince & Pinker are also concerned 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 (e.g., all-or-none processes) can emerge from processes that are soft and statistical. But it 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.

The rest of **Prince & Pinker**'s commentary seems to follow the same pattern: Here's something X that's easy for a symbolic model; X is hard for a connectionist model; look at the Rumelhart and McClelland model's first stab at trying to do X; that isn't good enough; therefore connectionist models can't possibly do X – in fact, "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 under way on how to do X better, and that the state of the art is already several years beyond what Prince & Pinker critique. The conclusion that connectionist models are "demonstrably inadequate," on the basis of the investigation of a single model representative of the previous generation of connectionist technology, seems grossly premature. As I state in no uncertain terms in (1), it currently seems quite unknowable whether connectionist models can adequately solve the problems they face.

The commentary of **Freidin** seems to be 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 then no doubt 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 connectionist learning system that has ever been built has, with greater or lesser success, solved 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 nontargets, i.e., 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 traditional point of view that a connectionist model, PARSNIP, that successfully learns to

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

model performance on grammaticality judgments, doesn't learn "grammar" (because Chomsky has patented the term to apply to something else). What is a bit puzzling is that less than a half dozen paragraphs earlier, Freidin claims 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, Mass., 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 embedded in complex symbolic rules, the innate endowment is a lot of continuous parameters embedded in simple numerical "rules." There is no *obvious* way in which the abstractions entering in the symbolic innate rules can be embedded in the innate structure of a connectionist network, but it is far too early to tell whether there is a nonobvious way that something equivalent can – and should – be done.

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 able to learn what they must learn? I believe that Shepard is right, both in characterizing this as a 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 et al.** raise the same issue of grappling with the prior commitments embedded in network architectures.)

Can patterns of activity, **Rey** wonders, 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 of 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 [personal communication], that motivated this research.)

The commentary of **Lehnert** makes the following points:

(a) Psychologists are attracted to connectionism because of theorem envy.

(b) Connectionism is methodology-driven research and that's dangerous.

(c) The methodology driving Smolensky is physics.

(d) Smolensky wants to ignore representational issues.

The presupposition of (a) is 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 happens that the school of symbolic AI to which Lehnert belongs prefers to regard such theorems as irrelevant.

Point (b) is not argued, simply asserted. It does not seem correct, but for the sake of argument, I will accept it. 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 methodology-driven research, and researchers who now want to apply it can afford the luxury of focusing exclusively on problem domains. The connectionist community as a whole cannot afford that luxury at this time. Some of us have to worry about the methodological problems, and we each apply the tools from which we think we can get the most mileage. Mine happen to be tools from physics.

I do not understand why **Lehnert** thinks I believe representational issues are not central to connectionism. 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 (e.g. a linguistic task) and a connectionist solution." This quote is from Smolensky (1987, p. 1), 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 lists and trees. This paper's results are presented, unfortunately for Lehnert, as theorems, but the work should leave little doubt about the importance I attach to issues of connectionist representation.

First it is claimed by **Hunter** that 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-inconsistently)[6] defined set of networks; he 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 would in fact, seem best taken to refer to exactly the systems I defined them to refer to, and not to the small set Hunter considers. Whatever weaknesses the target article may have, extreme narrowness of the framework is not among them.

The comments of **McCarthy** about unary fixation and lack of "elaboration tolerance" seem to be on target. At this point, connectionist models tend to be developed with the barest minimum of representational power for the target task. If the task is beefed up even a little, 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 – the possibility of doing more is not usually entertained.

It would be a mistake to leave the impression that

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

connectionist models cannot represent relations higher than unary. 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 and Kawamoto's (1986) work on semantic role assignment, Derthick's (1986; 1987) connectionist implementation of a micro-KL-ONE, Touretzky and Hinton's (1985) connectionist implementation of a simple production system, and my work (Smolensky 1987) on the representation of symbolic structures. In much of this work, the key is to use microfeatures that denote the conjunction of a microfeature 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 to 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 pronunciation. 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 pronunciation 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 settle quickly 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 the 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* and *X*, letters were used that are pronounceable in English in a larger variety of contexts (e.g., 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 pronunciation of *Q* and *X;* these connections are established only slowly through practice. But at instruction time, it *is* necessary to change instantly 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 ability to understand a language requires a network that has many "virtual harmony maxima" at states corresponding to the well-formed sentences and their meanings (what I call "virtual memories"); 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.

ACKNOWLEDGMENTS

This research has been supported by NSF grants IRI-8609599 and ECE-8617947 by the Sloan Foundation's computational neuroscience program, and by the Department of Computer Science and Institute of Cognitive Science at the University of Colorado at Boulder.

NOTES

1. Some commentators use "implementation" loosely, apparently equating it with a weaker notion such as instantiation (e.g., **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 "merely implementational" refer to this specific sense of "implementation."

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

3. Note that just the reverse is true of implementations: There, according to the microaccount, the ontology of the macroaccount *must* exist, since it can be logically and exactly derived from the microaccount. 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 the comments of **Cleland, Woodfield & Morton,** and **Prince & Pinker.** Given this, the hypothesis in (10) of the target article (that the subsymbolic account is complete) should not be construed to refer to *explanatory* completeness but rather to the sense of "completeness" in which quantum theory *in principle* applies to all phenomena, whereas Newtonian mechanics does not.

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

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

References

Ackley, D. H., Hinton, G. E. & Sejnowski, T. J. (1985) A learning algorithm for Boltzmann machines. *Cognitive Science* 9:147–69.

Allen, R. (1987) Natural language and back-propagation: Demonstratives, analogies, pronoun reference, and translation. *Proceedings of the IEEE First Annual International Conference on Neural Networks,* San Diego, Calif.

Alvarado, S., Dyer, M. G. & Flowers, M. (1986) Editorial comprehension of OpEd through argument units. *Proceedings of the American Association of Artificial Intelligence (AAAI-86),* Philadelphia, Pa.

Anderson, D. Z. (1986) Coherent optical eigenstate memory. *Optics Letters* 11:56–58.

Anderson, J. A. & Mozer, M. C. (1981) Categorization and selective neurons. In: *Parallel models of associative memory,* ed. G. E. Hinton & J. A. Anderson. Erlbaum.

Anderson, J. A., Silverstein, J. W. & Ritz, S. A. (1977) Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review* 84:413–51.

Anderson, J. R. (1981) *Cognitive skills and their acquisition.* Erlbaum.

(1983) *The architecture of cognition.* Harvard University Press.

(1985) *Cognitive Science* 9(1): Special issue on connectionist models and their applications.

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

Ashby, W. R. (1952) *Design for a brain*. Chapman and Hall.
Baird, B. (1986) Nonlinear dynamics of pattern formation and pattern recognition in the rabbit olfactory bulb. *Physica* 22D:150–75.
Ballard, D. H. (1986) Cortical connections and parallel processing: Structure and function. *Behavioral and Brain Sciences* 9:67–120.
Ballard, D. H. & Hayes, P. J. (1984) Parallel logical inference. *Proceedings of the Sixth Conference of the Cognitive Science Society*.
Baron, R. J. (1987) *The cerebral computer: An introduction to the computational structure of the human brain*. Erlbaum.
Barsalou, L. (1986) The instability of graded structure: Implications for the nature of concept. In: *Concepts and conceptual development: Ecological and intellectual factors in categorization*, ed. U. Neisser. Cambridge University Press.
(in press) Intra-concept similarity and its implications for inter-concept similarity. In: *Similarity and analogical reasoning*, ed. S. Vosniadou & A. Ortony. Cambridge University Press.
Barwise, J. (1986) Information and circumstance. *Notre Dame Journal of Formal Logic* 27(3):324–38.
Bechtel, W. (1985) Realism, instrumentalism, and the intentional stance. *Cognitive Science* 9:473–97.
(1988) *Philosophy of science: An overview for cognitive science*. Erlbaum.
Belew, R. K. (1986) Adaptive information retrieval: Machine learning in associative networks. Ph.D. thesis, Computer Science Department, University of Michigan.
Birnbaum, L. A. (1986) *Integrated processing in planning and understanding*. Ph.D. Thesis, Yale University.
Burge, T. (1986) Individualism and psychology. *Philosophical Review* 95(1):3–45.
Chandrasekaran, B. (1987) Towards a functional architecture for intelligence based on generic information processing tasks. *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, Milan, Italy.
(1988) What kind of information processing is intelligence? A perspective on AI paradigms and a proposal. *Foundations of AI: A Source Book*, ed. Partridge & Wilks. Cambridge University Press.
Charniak, E. (1987) Connectionism and explanation. In: *Proceedings of Theoretical Issues in Natural Language Processing* 3:68–72. New Mexico State University.
Chomsky, N. (1972) *Language and mind*. Harcourt.
(1980a) *Rules and representations*. Columbia University Press.
(1980b) *Rules and representations. Behavioral and Brain Sciences* 3:1–61.
(1984) *Lectures on government and binding* (2nd rev.). Foris.
Churchland, P. S. (1984) *Matter and consciousness*. MIT Press.
(1986) *Neurophilosophy*. MIT Press/Bradford Books.
Cohen, M. S. (1986) Design of a new medium for volume holographic information processing. *Applied Optics* 14:2288–94.
Cooper, L. A. (1976) Demonstration of a mental analog of an external rotation. *Perception & Psychophysics* 19:296–302.
Cottrell, G. (1987) Toward connectionist semantics. In: *Proceedings of Theoretical Issues in Natural Language Processing* 3:63–67. New Mexico State University.
Cox, R. T. (1946) Probability, frequency, and reasonable expectation. *American Journal of Statistical Physics* 14:1–13.
Crick, F. & Asanuma, C. (1986) Certain aspects of the anatomy and physiology of the cerebral cortex. In: *Parallel distributed processing: Explorations in the microstructure of cognition*. Vol. 2: *Psychological and biological models*, ed. J. L. McClelland, D. E. Rumelhart & the PDP Research Group. MIT Press/Bradford Books.
Davidson, D. (1970) Mental events. In: *Experience and theory*, ed. L. Foster & J. W. Swanson. University of Massachusetts Press.
DeJong, G. (1983) An approach to learning from observation. *Proceedings of the International Machine Learning Workshop*. University of Illinois.
Dell, G. S. (1985) Positive feedback in hierarchical connectionist models: Applications to language production. *Cognitive Science* 9:3–23.
Dennett, D. C. (1986) The logical geography of computational approaches: A view from the east pole. In: *The representation of knowledge and belief*, ed. M. Brand & R. M. Harnish. University of Arizona Press.
Derthick, M. A. (1986) *A connectionist knowledge representation system*. Thesis proposal, Computer Science Department, Carnegie-Mellon University.
(1987a) Counterfactual reasoning with direct models. *Proceedings of the Sixth National Conference on Artificial Intelligence*, Seattle, Wash., 346–51.
(1987b) A connectionist architecture for representing and reasoning about structured knowledge. *Proceedings of the Ninth Annual Conference of the Cognitive Science Society*, Seattle, Wash., 131–42.
Dolan, C. & Dyer, M. G. (1987) Evolution of an architecture for symbol processing. *Proceedings of the IEEE First Annual International Conference on Neural Networks*, San Diego, Calif.
(1987) Symbolic schemata, role binding, and the evolution of structure in connectionist memories. *Proceedings of the IEEE First Annual International Conference on Neural Networks*, San Diego, Calif.
Dresher, E. & Hornstein, N. (1976) On some supposed contributions of artificial intelligence to the scientific study of language. *Cognition* 4:321–98.
Dreyfus, H. L. & Dreyfus, S. E. (in press) Making a mind vs. modeling the brain: AI back at a branchpoint. *Daedalus*.
Feigenbaum, E. A. (1963) The simulation of verbal learning behavior. In: *Computers and thought*, ed. E. A. Feigenbaum & J. Feldman. McGraw-Hill.
Feldman, J. A. (1981) A connectionist model of visual memory. In: *Parallel models of associative memory*, ed. G. E. Hinton & J. A. Anderson. Erlbaum.
(1985) Four frames suffice: A provisional model of vision and space. *Behavioral and Brain Sciences* 8:265–89.
(1986) Neural representation of conceptual knowledge. Technical Report 189, Department of Computer Science, University of Rochester.
Feldman, J. A. & Ballard, D. H. (1982) Connectionist models and their properties. *Cognitive Science* 6:205–54.
Feldman, J. A., Ballard, D. H., Brown, C. M. & Dell, G. S. (1985) Rochester connectionist papers: 1979–1985. Technical Report 172, Department of Computer Science, University of Rochester.
Fields, C. & Dietrich, E. (1987) A stochastic computing architecture for multi-domain problem solving. *Proceeding of the International Symposium on Methodologies for Intelligent Systems* (in press).
Fodor, J. A. (1975) *The language of thought*. Harvard University Press.
(1986) Information and association. *Notre Dame Journal of Formal Logic* 27:307–23.
(1987) Why there still has to be a language of thought. In: *Psychosemantics*, ed. J. A. Fodor. MIT Press/Bradford Books.
(1983) *The modularity of mind*. MIT Press.
Fodor, J. A. & Pylyshyn, Z. W. (1988) Connectionism & cognitive architecture: A critical analysis. *Cognition* 28: 3–71.
Freeman, W. J. (1975) *Mass action in the nervous system*. Academic Press.
(1987) Simulation of chaotic EEG patterns with a dynamic model of the olfactory system. *Biological Cybernetics* 56:139–50.
(in press) Hardware simulation of brain dynamics of learning: The SPOCK. *First International IEEE Conference on Neural Networks*, San Diego.
Freeman, W. J. & Skarda, C. A. (1985) Spatial EEG patterns, nonlinear dynamics and perception: The neo-Sherringtonian view. *Brain Research Reviews* 10:147–75.
Freidin, R. (1987) *Foundations of generative syntax*. Manuscript.
Geman, S. & Geman, D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6:721–41.
Gillispie, C. C. (1960) *The edge of objectivity: An essay in the history of scientific ideas*. Princeton University Press.
Golden, R. M. (submitted) A probabilistic computational framework for neural network models.
Goldman, A. I. (1970) *A theory of human action*. Prentice-Hall.
Grossberg, S. (1976) *Adaptive pattern classification and universal recoding. Biological Cybernetics* 23:121–34; 187–202 (in two parts).
(1980) How does the brain build a cognitive code? *Psychological Review* 87:1–51.
(1987) Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science* 11:23–63.
ed. (1987a) *The adaptive brain I: Cognition, learning, reinforcement, and rhythm*. North-Holland.
ed. (1987b) *The adaptive brain II: Vision, speech, language, and motor control*. North-Holland.
Grossberg, S. & Mingolla, E. (1985) Neural dynamics of form perception: Boundary completion, illusory figures, and neon color spreading. *Psychological Review* 92:173–211.
Grossberg, S. & Stone, G. (1986) Neural dynamics of word recognition and recall: Attentional priming, learning, and resonance. *Psychological Review* 93:46–74.
(1986) Neural dynamics of attention switching and temporal order information in short-term memory. *Memory and Cognition* 14:451–68.
Halle, M. (1962) Phonology in generative grammar. *Word* 18:54–72.
Hammond, K. J. (1986) *Case-based planning: An integrated theory of planning, learning and memory*. Ph.D. Thesis, Yale University.
Hanson, S. J. & Kegl, J. (1987) PARSNIP: A connectionist network that learns natural language grammar on exposure to natural language sentences. Unpublished manuscript, Bell Communications Research and Princeton University.
Haugeland, J. (1978) The nature and plausibility of cognitivism. *Behavioral and Brain Sciences* 1:215–26.

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

Hebb, D. O. (1949) *The organization of behavior.* Wiley.

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

Hinton, G. E. & Anderson, J. A., eds. (1981) *Parallel models of associative memory.* Erlbaum.

Hinton, G. E., McClelland, J. L. & Rumelhart, D. E. (1986) Distributed representations. In: *Parallel distributed processing: Explorations in the microstructure of cognition.* Vol. 1: *Foundations*, ed. J. L. McClelland, D. E. Rumelhart & the PDP Research Group. MIT Press/Bradford Books.

Hinton, G. E. & Plaut, D. C. (1987) Using weights to deblur old memories. In: *Proceedings of the Ninth Annual Cognitive Science Society Conference.* Erlbaum.

Hinton, G. E. & Sejnowski, T. J. (1983a) Analyzing cooperative computation. *Proceedings of the Fifth Annual Conference of the Cognitive Science Society.*

(1983b) Optimal perceptual inference. *Proceedings of the I.E.E.E. Conference on Computer Vision and Pattern Recognition.*

(1986) Learning and relearning in Boltzmann machines. In: *Parallel distributed processing: Explorations in the microstructure of cognition.* Vol. 1: *Foundations*, ed. J. L. McClelland, D. E. Rumelhart & the PDP Research Group. MIT Press/Bradford Books.

Hinton, G. E., Sejnowski, T. J. & Ackley, D. H. (1984) Boltzmann machines: Constraint satisfaction networks that learn. Technical report CMU-CS-84-119, Computer Science Department, Carnegie-Mellon University.

Hofstadter, D. R. (1979) *Godel, Escher, Bach: An eternal golden braid.* Basic Books.

(1983) The architecture of Jumbo. *Proceedings of the International Machine Learning Workshop.*

(1984) The Copycat Project: An experiment in nondeterminism and creative analogies. AI Memo #755, MIT Artificial Intelligence Laboratory.

(1985a) Variations on a theme as the crux of creativity. In: *Metamagical themas.* Basic Books.

(1985b) Waking up from the Boolean dream, or, subcognition as computation. In: *Metamagical themas.* Basic Books.

Holland, J. H. (1986) Escaping brittleness: The possibilities of general-purpose machine-learning algorithms applied to parallel rule-based systems. In: *Machine learning: An artificial-intelligence approach*, vol. 2, ed. R. S. Michalski et al. William Kaufmann.

Holland, J. H., Holyoak, K. J., Nisbitt, R. E. & Thagard, P. (1986) Induction: Processes of learning, inference and discovery. Bradford Books.

Hopcroft, J. E. & Ullman, J. D. (1979) *Introduction to automata theory, languages, and computation.* Addison-Wesley.

Hopfield, J. J. (1982) *Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Science* 79:2554–58.

(1984) Neurons with graded response have collective properties like those of two-state neurons. *Proceedings of the National Academy of Sciences, USA* 81:3088–92.

Horgan, T. (1982) Supervenience and microphysics. *Pacific Philosophical Quarterly* 63:29–43.

Hummel, R. A. & Zucker, S. W. (1983) On the foundations of relaxation labeling processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 5:267–87.

Jakobson, R., Fant, G. & Halle, M. (1951) *Preliminaries to speech analysis.* MIT Press.

James, W. (1967) *The writings of William James: A comprehensive edition.* Random House.

Jeffreys, H. (1983) *Theory of probability.* Clarendon.

Jordan, M. I. (1986) Attractor dynamics and parallelism in a connectionist sequential machine. *Proceedings of the Eighth Meeting of the Cognitive Science Society.*

Kaas, A. (1986) Modifying explanations to understand stories. *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*, Amherst, Ma.

Kant, I. (1787/1963) *Critique of pure reason.* N. Kemp Smith, trans., 2nd ed. Macmillan.

Kohonen, T. (1984) *Self-organization and associative memory.* Springer-Verlag.

Kosslyn, S. M. (1987) Seeing and imaging in the cerebral hemispheres: A computational approach. *Psychological Review* 94:148–75.

Kripke, S. (1980) *Wittgenstein on following a rule.* Harvard University Press.

Lakoff, G. (1987) *Women, fire, and dangerous things: What categories reveal about the mind.* University of Chicago Press.

Langacker, R. (1987) *Foundations of cognitive grammar.* Stanford University Press.

Larkin, J. H., McDermott, J., Simon, D. P. & Simon, H. A. (1980) Models of competence in solving physics problems. *Cognitive Science* 4:317–45.

Lashley, K. (1950) In search of the engram. In: *Psychological mechanisms in animal behavior*, Symposia of the Society for Experimental Biology, No. 4. Academic Press.

Lewis, C. H. (1978) *Production system models of practice effects.* Unpublished doctoral dissertation, University of Michigan.

Luenberger, D. G. (1984) *Linear and nonlinear programming.* Addison-Wesley.

Luria, A. R. (1966) *Higher cortical functions in man.* Basic Books.

Lycan, W. G. (1981) Form, function, and feel. *Journal of Philosophy* 78:24–50.

(1987) *Consciousness.* MIT Press/Bradford Books.

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

McClelland, J. L. (1987) Parallel distributed processing and role assignment constraints. In: *Proceedings of Theoretical Issues in Natural Language Processing* 3:73–77. New Mexico State University.

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

McClelland, J. L. & Rumelhart, D. E. (1981) An interactive activation model of context effects in letter perception: Part 1. An account of the basic findings. *Psychological Review* 88:375–407.

McClelland, J. L., Rumelhart, D. E. & the PDP Research Group (1986) *Parallel distributed processing: Explorations in the microstructure of cognition.* Vol. 2: *Psychological and biological models.* MIT Press/Bradford Books.

Miikkulainen, R. & Dyer, M. G. (1987) Building distributed representations without microfeatures. Technical report UCLA-AI-87-17, Computer Science Dept., UCLA, Los Angeles, Calif.

Miller, G. A. & Chomsky, N. (1963) Finitary models of language users. In: *Handbook of mathematical psychology*, ed. R. R. Bush, E. Galanter & R. D. Luce. Wiley.

Minsky, M. (1963) Steps toward artificial intelligence. In: *Computers and thought*, ed. E. A. Feigenbaum & J. Feldman. McGraw-Hill.

(1975) A framework for representing knowledge. In: *The psychology of computer vision*, ed. P. H. Winston. McGraw-Hill.

Mishkin, M., Malamut, B. & Bachevalier, J. (1984) Memories and habits: Two neural systems. In: *Neurobiology of learning and memory*, ed. G. Lynch, J. L. McGauth & N. M. Weinberger.

Minsky, M. & Papert, S. (1969) *Perception.* MIT Press.

Moran, J. & Desimone, R. (1985) Selective attention gates visual processing in the extrastriate cortex. *Science* 229:782–84.

Nelson, R. J. (1982) *The logic of mind.* D. Reidel.

(1987a) Church's Thesis in cognitive science. *Notre Dame Journal of Formal Logic.* Forthcoming.

(1987b) Models for cognitive science. *Philosophy of Science.* Forthcoming.

Newell, A. (1980) Physical symbol systems. *Cognitive Science* 4:135–83.

Newell, A., Shaw, J. C. & Simon, H. A. (1958) Elements of a theory of human problem solving. *Psychological Review* 65:151–66.

Newell, A. & Simon, H. A. (1972) *Human problem solving.* Prentice-Hall.

(1976) Computer science as empirical inquiry: Symbols and search. *Communications of the Association for Computing Machinery* 19:113–26.

Oden, G. C. (1977) Integration of fuzzy logical information. *Journal of Experimental Psychology: Human Perception and Performance* 3:565–75.

(1987) Concept, knowledge, and thought. *Annual Review of Psychology* 38:203–28.

Pazanni, M. & Dyer, M. G. (1987) A comparison of concept identification in human learning and network learning with the generalized delta rule. *Proceedings of the Tenth International Joint Conference on Artificial Intelligence (IJCAI-87).* August, Milan, Italy.

Pearl, J. (1985) Bayesian networks: A model of self-activated memory for evidential reasoning. *Proceedings of the Seventh Conference of the Cognitive Science Society.*

Pinker, S. (1984) *Language learnability and language development.* Harvard University Press.

Pinker, S. & Prince, A. (1988) On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition* 28: 73–193.

Pollack, J. B. (1987) On connectionist models of natural language processing MCCS-87-100. Ph.D. dissertation, Computing Research Laboratory, New Mexico State University.

Putnam, H. (1975) Philosophy and our mental life. In: *Philosophical papers*, vol. 2. Cambridge University Press.

(1980) Philosophy and our mental life. In: *Readings in philosophy of psychology*, vol. 1, ed. N. Block. Harvard University Press.

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

Pylyshyn, Z. W. (1984) *Computation and cognition: Toward a foundation for cognitive science*. MIT Press/Bradford Books.

Reddy, D. R., Erman, L. D., Fennell, R. D. & Neely, R. B. (1973) The HEARSAY speech understanding system: An example of the recognition process. *Proceedings of the Third International Joint Conference on Artificial Intelligence*, Stanford, Calif.

Reeves, A. & Sperling, G. (1986) Attentional theory of order information in short-term memory. *Psychological Review* 93:180–206.

Rey, G. (1983) Concepts and stereotypes. *Cognition* 15:237–62.

(1985) Concepts and conceptions. *Cognition* 19:297–303.

Riley, M. S. & Smolensky, P. (1984) A parallel model of (sequential) problem solving. *Proceedings of the Sixth Annual Conference of the Cognitive Science Society*.

Rumelhart, D. E. (1975) Notes on a schema for stories. In: *Representation and understanding*, ed. D. G. Bobrow & A. Collins. Academic Press.

(1980) Schemata: The building blocks of cognition. In: *Theoretical issues in reading comprehension*, ed. R. Spiro, B. Bruce & W. Brewer. Erlbaum.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986) Learning and internal representations by error propagation. In: *Parallel distributed processing: Explorations in the microstructure of cognition*. Vol. 1: *Foundations*, ed. J. L. McClelland, D. E. Rumelhart & the PDP Research Group. MIT Press/Bradford Books.

Rumelhart, D. E. & McClelland, J. L. (1982) An interactive activation model of context effects in letter perception: Part 2. The contextual enhancement effect and some tests and extensions of the model. *Psychological Review* 89:60–94.

(1986) On learning the past tenses of English verbs. In: *Parallel distributed processing: Explorations in the microstructure of cognition*. Vol. 2: *Psychological and biological models*, ed. J. L. McClelland, D. E. Rumelhart & the PDP Research Group. MIT Press/Bradford Books.

(1986a) PDP models and general issues in cognitive science. In: *Parallel distributed processing: Explorations in the microstructure of cognition*, vol. 1, ed. J. L. McClelland, D. E. Rumelhart & the PDP Research Group. MIT Press/Bradford Books.

Rumelhart, D. E., McClelland, J. L. & the PDP Research Group (1986) *Parallel distributed processing: Explorations in the microstructure of cognition*. Vol. 1: *Foundations*. MIT Press/Bradford Books.

Rumelhart, D. E., Smolensky, P., McClelland, J. L. & Hinton, G. E. (1986) Schemata and sequential thought processes in parallel distributed processing models. In: *Parallel distributed processing: Explorations in the microstructure of cognition*. Vol. 2: *Psychological and biological models*, ed. J. L. McClelland, D. E. Rumelhart & the PDP Research Group. MIT Press/Bradford Books.

Savage, L. J. (1971) *The foundations of statistics*. Wiley.

Schank, R. C. (1972) Conceptual dependency: A theory of natural language understanding. *Cognitive Psychology* 3(4):552–631.

(1975) *Conceptual information processing*. North-Holland.

(1982) *Dynamic memory: A theory of learning in computers and people*. Cambridge University Press.

Schank, R. C. & Abelson, R. (1977) *Scripts, plans, goals and understanding*. Erlbaum.

Schank, R. C., Collins, G. C. & Hunter, L. E. (1986) Transcending inductive category formation in learning. *Behavioral and Brain Sciences* 9:639–86.

Schneider, W. & Detweiler, M. (1987) A connectionist/control architecture for working memory. In: *The psychology of learning and motivation*, vol. 21, ed. G. H. Bower. Academic Press.

Schneider, W. & Mumme, D. (forthcoming) Attention, automaticity and the capturing of knowledge: A two-level cognitive architecture.

Searle, J. R. (1980) Minds, brains, and programs. *Behavioral and Brain Sciences* 3:417–57.

Sejnowski, T. J. (1976) On the stochastic dynamics of neuronal interactions. *Biological Cybernetics* 22:203–11.

Sejnowski, T. J. & Rosenberg, C. R. (1986) NETtalk: A parallel network that learns to read aloud. Technical Report JHU/EECS-86/01, Department of Electrical Engineering and Computer Science, John Hopkins University.

(1987) Parallel networks that learn to pronounce English text. *Complex Systems* 1:145–68.

Selfridge, O. G. (1959) Pandemonium: A paradigm for learning. In: *Symposium on the mechanisation of thought processes*. London: H. M. Stationery Office.

Shastri, L. (1985) Evidential reasoning in semantic networks: A formal theory and its parallel implementation. Technical Report TR 166, Department of Computer Science, University of Rochester.

(1987) A connectionist encoding of semantic networks. In: *Proceedings of the Ninth Annual Conference of the Cognitive Science Society*, 143–54.

Shastri, L. & Feldman, J. A. (1985) Evidential reasoning in semantic networks: A formal theory. In: *Proceedings of the International Joint Conference on Artificial Intelligence*, Los Angeles, 465–74.

(1988)The constituent structure of connectionist mental states: A reply to Fodor and Pylyshyn, *Southern Journal of Philosophy*. Special issue on connectionism and the foundations of cognitive science. In press.

Shepard, R. N. (1962) The analysis of proximities: Multidimensional scaling with an unknown distance function. I & II. *Psychometrika* 27:125–40, 219–46.

(1964) Review of *Computers and thought* (ed. E. Feigenbaum & J. Feldman). *Behavioral Science* 9:57–65.

(1981) Psychophysical complementarity. In: *Perceptual organization*, ed. M. Kubovy & J. Pomerantz. Erlbaum.

(1984) Ecological constraints on internal representation: Resonant kinematics of perceiving, imagining, thinking, and dreaming. *Psychological Review* 91:417–47.

(1987) Towards a universal law of generalization for psychological science. *Science* (in press).

Shepard, R. N. & Cooper, L. A. (1982) *Mental images and their transformations*. MIT Press/Bradford Books.

Shepard, R. N. & Metzler, J. (1971) Mental rotation of three-dimensional objects. *Science* 171:701–3.

Shiffrin, R. M. & Schneider, W. (1977) Controlled and automatic human information processing. II. Perceptual learning. *Psychological Review* 84:127–90.

Skarda, C. A. & Freeman, W. J. (1987) How brains make chaos in order to make sense of the world. *Behavioral and Brain Sciences* 10:161–95.

Smale, S. (1987) On the topology of algorithms, I. *Journal of Complexity* 3: 81–89.

Smart, J. J. C. (1959) Sensations and brain processes. *Philosophical Review* 141–56.

Smolensky, P. (1983) Schema selection and stochastic inference in modular environments. *Proceedings of the National Conference on Artificial Intelligence*.

(1984a) Harmony theory: Thermal parallel models in a computational context. In: *Harmony theory: Problem solving, parallel cognitive models, and thermal physics*, ed. P. Smolensky & M. S. Riley. Technical Report 8404, Institute for Cognitive Science, University of California at San Diego.

(1984b) The mathematical role of self-consistency in parallel computation. *Proceedings of the Sixth Annual Conference of the Cognitive Science Society*.

(1986a) Information processing in dynamical systems: Foundations of harmony theory. In: *Parallel distributed processing: Explorations in the microstructure of cognition*. Vol. 1: *Foundation*, ed. J. L. McClelland, D. E. Rumelhart & the PDP Research Group. MIT Press/Bradford Books.

(1986b) Neural and conceptual interpretations of parallel distributed processing models. In: *Parallel distributed processing: Explorations in the microstructure of cognition*. Vol. 2: *Psychological and biological models*, ed. J. L. McClelland, D. E. Rumelhart & the PDP Research Group. MIT Press/Bradford Books.

(1986c) Formal modeling of subsymbolic processes: An introduction to harmony theory. In: *Directions in the science of cognition*, ed. N. E. Sharkey. Ellis Horwood.

(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.

(1987a) Connectionist AI, symbolic AI, and the brain. *Artificial Intelligence Review* 1:95–109.

(1988) The constituent structure of connectionist mental states: A reply to Fodor and Pylyshyn. *Southern Journal of Philosophy*. Special issue on connectionism and the foundations of cognitive science.

Stich, S. P. (1983) *From folk psychology to cognitive science*. MIT Press.

Tolman, E. C. (1932) *Purposive behavior in animals and men*. Appleton-Century-Crofts.

Toulouse, G., Dehaene, S. & Changeux, J.-P. (1986) A spin glass model of learning by selection. Technical Report, Unite de Neurobiologie Moleculaire, Institut Pasteur, Paris.

Touretzky, D. S. (1986) BoltzCONS: Reconciling connectionism with the recursive nature of stacks and trees. *Proceedings of the Eighth Conference of the Cognitive Science Society*.

(1987) Representing conceptual structures in a neural network. *Proceedings of the IEEE First Annual International Conference on Neural Networks*, San Diego, Calif.

Touretzky, D. S. & Geva, S. (1987) A distributed connectionist representation for concept structures. In: *Proceedings of the Ninth Annual Conference of the Cognitive Science Society*, 155–64.

Touretzky, D. S. & Hinton, G. E. (1985) Symbols among the neurons: Details of a connectionist inference architecture. *Proceedings of the International Joint Conference on Artificial Intelligence*.

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press

Turing, A. (1936) On computable numbers, with an application to Entscheidungs problem. *Proceedings of the London Mathematical Society* (Ser. 2) 42:230–65 and 43:544–46.

Tversky, A. & Kahneman, D. (1983) Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological Review* 90:293–315.

Ullman, S. (1984) Visual routines. *Cognition* 18:97–159.

den Uyl, M. J. (1986) Representing magnitude by memory resonance: A hypothesis on qualitative judgment. *Proceedings of the Eighth Annual Conference on the Cognitive Science Society.*

Van Essen, D. C. (1985) Functional organization of primate visual cortex. *The cerebral cortex,* vol. 3.

Waldrop, M. M. (1984) Artificial intelligence in parallel. *Science* 225:608–10.

Waltz, D. L. (1978) An English language question answering system for a large relational database. *Communications of the Association for Computing Machinery* 21:526–39.

Waltz, D. L. & Pollack, J. B. (1985) Massively parallel parsing: A strongly interactive model of natural language interpretation. *Cognitive Science* 9:51–74.

Wilks, Y. A. (1978) Making preference more active. *Artificial Intelligence* 11:197–223.

Wittgenstein, L. (1956) *Remarks on the foundations of mathematics.* Blackwell.

https://doi.org/10.1017/S0140525X00052791 Published online by Cambridge University Press