Connectionism, Learning and Meaning

MORTEN H. CHRISTIANSEN & NICK CHATER

There is an apparent anomaly in the notion that connectionism, which is fundamentally a new technology, *has considerable* philosophical *significance. Nonetheless, connectionism has been widely viewed as having implications for symbol grounding, notions of structured representation and compositionality, as well as the issue of nativism. In this paper, we consider each of these issues in detail and find that the current state of connectionism does not warrant the magnitude of many of the philosophical conclusions drawn from it. We argue that connectionist models are no more 'grounded' than their classical counterparts. In addition, since connectionist representations typically are ascribed content through semantic interpretation based on correlation, connectionism is prone to a number of well known philosophical problems facing any kind of correlational semantics. However, we suggest that philosophy may be ill advised to ignore the development of connectionism, particularly if connectionist systems prove to be able to learn to handle structured representations.*

KEYWORDS: Compositionality, computation, connectionism, learning, nativism, representation, semantics.

1. Introduction

The surge of interest in neural networks has created an impact across a remarkably broad range of disciplines—from electrical engineering (Graf *et al.*, 1988) physics (Hopfield, 1982) and mathematics (Cybenko, 1989) to the biological and cognitive sciences (McClelland & Rumelhart, 1986; Rumelhart & McClelland, 1986), artificial intelligence (Derthick, 1987) and computer science (Fahlman, 1988). What may appear incongruous is that this remarkable influence has been widely considered to have significant *philosophical* implications too (Clark, 1989; Churchland, 1986; Horgan & Tienson, 1987; Churchland, 1989; Bechtel & Abrahamsen, 1991; Ramsey *et al.*, 1991). Indeed, it has been argued that neural networks have important ramifications for one of the most abstract areas of philosophy—the theory of meaning (Cottrell, 1987; Cussins, 1990; Harnad, 1990b, 1992).

At first blush, this state of affairs is surprising since, after all, neural networks are fundamentally a *technology*—a set of tools and methods which can be applied to a wide variety of practical and modelling tasks. Across the spread of disciplines caught

M. H. Christiansen, Centre for Cognitive Science, University of Edinburgh, 2 Buccleuch Place, Edinburgh EH8 9LW, UK. E-mail: morten@uk.ac.ed.cogsci. Nick Chater, Department of Psychology, University of Edinburgh, 7 George Square, Edinburgh EH8 9JZ, UK. E-mail: nicholas@uk.ac.ed.cogsci.

up in the study of neural networks most either apply formal methods well suited to describing and analyzing the behaviour of neural networks or are domains to which neural network modelling techniques can usefully be applied. Yet philosophy in general, and theory of meaning in particular, appear to stand in neither of these relations to neural networks—philosophical methods do not appear obviously helpful in elucidating network behaviour and it seems almost incoherent that neural networks could in some sense model or solve some philosophical problem.

In this paper, we argue that the purported philosophical implications of connectionism for the theory of meaning are, at least in many cases, illusory. Before doing so let us dwell briefly on why, despite this apparent poor fit, neural networks have been taken to have far reaching philosophical implications. The answer, of course, is that neural networks in their connectionist guise have been seen as providing a new *metaphor for the mind.* Of particular interest is the suggestion that connectionism is seen as providing a new account of the nature of *mental representation,* which is held to have a wide variety of philosophical ramifications. It is not always entirely clear which aspects of neural network computation are philosophically significant, although the fact that connectionist representations are typically *distributed* as well as *superpositional* and are usually *learnt* appears to be particularly important, as we shall see below. In any case, whatever the precise characteristics that do the philosophical work in connectionist modelling might be, connectionist models are interesting because they are *different:* different from the classical, symbolic view of cognitive processing which has dominated cognitive psychology and cognitive science since their inception (Fodor, 1975, 1987; Pylyshyn, 1984).[1]

However, it is still a controversial issue whether or not neural networks should be viewed as displacing symbolic accounts of mind or as a medium in which symbolic processes can be run (e.g. Clark, 1989; Fodor & Pylyshyn, 1988; Smolensky, 1988). Pinker & Prince (1988) have dubbed these positions *eliminativist connectionism* and *implementational connectionism,* respectively. At first sight, at least, it seems that neural networks will be of philosophical significance only for eliminative connectionists. It is hard to imagine how a new *implementation* could have any great philosophical implications at all. However, full scale eliminativism is a very radical position indeed and few in cognitive science would embrace it. There are, though, intermediate positions which allow symbolic representations and operations, but which still accord a connectionist substrate considerable importance in explaining cognition (e.g. Chater & Oaksford, 1990a; Clark, 1989; Harnad, 1990a). The arguments we consider for the philosophical significance of connectionism typically embrace both symbolic and connectionist explanations of cognition.

In discussions of symbolic and connectionist approaches to cognitive science, the historical predominance of the symbolic view has meant that, to some extent at least, the ground rules concerning what key cognitive phenomena must be explained and what counts as a good explanation, have been set in largely symbolic terms (van Gelder, 1992). Thus, if connectionism amounts to a genuinely new paradigm for the understanding of mind, there is a very real danger of falling into what we will call the '*incommensurability trap*'. That is, connectionist models may be unfairly judged either because they fail to fit the classical standards or because when they are made to fit the resulting explanation looks forced and unattractive. The danger is analogous to that of judging vegetarian food by the standards of the butcher. After all, connectionism—construed as a new paradigm (e.g. Schneider, 1987)—may involve a revolutionary reconstruction of the field from new fundamentals, leading

to changes in methodology and basic theoretical assumptions. Since rival paradigms prescribe different sets of standards and principles, connectionist and classical approaches to cognitive science may also differ on what constitute meaningful and legitimate scientific questions. Due to this incommensurability, discussions between proponents of different paradigms on the issue of paradigm choice often become circular. Each group will tend to praise their own and criticize the others party's paradigm with arguments based on their own paradigm. In other words, when comparing and assessing the individual explanatory power of rival paradigms, the incommensurability trap constitutes a non-trivial methodological obstacle to negotiate since it involves engaging in the process of *radical translation* (Quine, 1960). Or so much philosophy of science would have us to believe (e.g. Kuhn, 1970). In any case, there are signs that communication is becoming difficult and hence it is imperative that the merits of connectionism are judged from 'within', i.e. on its own terms, not through the looking glass of the classical paradigm.[2] However, the opposite danger is equally real—symbolic models can look unattractive from a connectionist perspective. This raises the danger of ignoring all that has been learnt from the symbolic approach and simply starting the project of understanding the mind afresh.

In particular, it is not possible to discuss the relationship between connectionism and the theory of meaning without flirting with these dangers, since traditional theory of meaning is concerned with symbolic, linguistic representations. So, for example, we shall break our discussion below into two parts—first discussing the semantics of primitive terms and then the issue of compositional semantics. This very distinction comes out of the theory of meaning for languages and need not necessarily be appropriate for other types of representation (if there are any). However, theories of meaning for languages are the only theories of meaning that we have—there is no alternative to which the connectionist can turn. In the main, this has meant that philosophical implications of connectionism have been viewed as fitting into a standard semantic framework: as either concerning fixing the meaning of primitives, or compositional semantics.

So far, we have considered the possible importance of connectionism for semantics. Equally interesting is the significance of semantic issues for connectionism itself. Ascribing meaning to connectionist networks involves implicitly making assumptions about what it is for a state of a network to represent. Without a theory of meaning, whether explicit or implicit, it is impossible to view networks as possessing or developing *representations* at all. More generally, seeing a connectionist network, or any other system, as a *computer* at all, is dependent on being able to ascribe meaning to the states of the system. Otherwise its internal states are not appropriately viewed as *processing information* at all, but simply as passing through sequence of states; the network will be viewed simply as an informational 'black box', where only inputs and outputs are interpreted, and those by fiat. Hence, the semantics of connectionist networks which we will discuss extensively below is of both practical as well as philosophical interest.

The structure of the paper is as follows. In Section 2, we give a brief exposition of the classical approach to cognitive processing in which the main object of a theory of meaning is to elucidate the semantic content of the internal language of thought. One of the largest problems facing this approach is the problem of establishing the right referential links between internal representations and the external world. Much optimism has been invested in connectionism as providing the means for referential grounding of semantic primitives through learning. We therefore address some of

the philosophical problems standing in the way of this project in Section 3, specifically the fact that connectionist representations, at least presently, are no more grounded than their symbolic counterpart (basically because they are developed from pre-processed input), and, more generally, that obtaining the correct correlations between internal representations and external objects is a non-trivial matter (i.e. it involves problems concerning error, underdetermination, non-existing entities, and the difference between properties and propositions). Whether or not connectionism will be able to ground semantic primitives, it does need to develop some kind of compositional semantics, if it is to be a true rival to the symbolic paradigm. Consequently, Section 4 discussed the issue of learning complex semantic representations in connectionist models. In particular, we outline what kind of compositionality we should envisage and point to initial steps taken in the direction of truly structure sensitive manipulations of connectionist representations. Since learning is one of the leading motivations behind connectionist modelling, we devote the last section to a discussion of issue of nativism in relation to symbolic as well as connectionist models. Specifically, we find that connectionism may provide the prospect of a better explanation of cognitive development.

2. The Classical View of Computation and Cognition

The classical view, baldly stated, is that cognitive processes are defined over sentences of an internal language in virtue of their form (Fodor, 1981). In particular, the classical model of cognition rests on two major claims. Firstly, psychological explanation is best carried out in terms of an internal language of thought. Second, this internal language involves a machine-implementable physical symbol system (Newell & Simon, 1976) with structure sensitive transformations of symbolic expressions on the level of syntax, i.e. classical representation of mental states can be implemented on computers. The classical paradigm consists of the synthesis of these two claims.

> If, in principle, syntactic relations can be made to parallel semantic relations, and if, in principle, you can have a mechanism whose operations on formulas are sensitive to their syntax, then it may be possible to construct a *syntactically* driven machine whose state transitions satisfy *semantical* criteria of coherence. Such a machine would be just what's required for a mechanical model of the semantical coherence of thought; correspondingly, the idea that the brain is such a machine is the foundational hypothesis of classical cognitive science. (Fodor & Pylyshyn, 1988, p. 30.)

They sum up this position in the slogan that cognition is mechanized proof theory. Actually, this position is rather stronger than the view that cognition is symbol manipulation, since it is by no means always appropriate to view symbol manipulation as a theorem proving (Chater & Oaksford, 1990b; Oaksford & Chater, 1991).

Like natural and logical languages, the internal language is assumed to consist of a finite stock of atomic primitives and a finite set of ways of combining these primitives (Dowty *et* al., 1981). These modes of combination can be applied arbitrarily often, to give an infinite set of possible internal formulae. Specifying a semantics for such a language involves specifying (i) the meanings of the primitives of the language and (ii) a compositional semantics for that language, which specifies how the meanings of the parts of a complex expression contribute to the meaning of the whole, given each possible mode of combination. Below we shall see that connectionism has been viewed as potentially impacting on both of these aspects of semantics.

What have been taken as the philosophical implications of the classical position? The most direct impact has been that the classical view allows (although it by no means requires) that the contents of internal formulae may be identified with the contents of mental states: propositional attitudes usually viewed as computational relations to mental representations (Fodor, 1975; Field, 1978). So, for example, the belief, desire or hope that P, for a proposition P, amounts to standing in the appropriate relation to an internal formula which expresses P. Where mental states are explained in terms of *propositions* represented by internal *sentences*, concepts are explained in terms of the properties represented by internal *predicates*. So, to have the concept DOG, DOG-WITH-ONE-LEG or whatever, is to possess an internal formula which expresses the properties of being a dog or being a one-legged dog. This position is attractively parsimonious in ontological terms: while there appear *prima facie* to be two sorts of entities with semantic properties, languages and mental states, there are, at root, only one, since the semantic properties of the latter are derivative on the former.

According to this picture, an important concern of the theory of meaning is to explain the basis for the semantic properties of internal languages. The project has a rather different character to the project of explaining the basis of the semantics of external natural language. It must be conducted in the absense of any detailed understanding of the nature of this language, and the social and conventional aspects of meaning in natural language appear to be relevant to the semantics of an internal language. Within philosophy there has been extensive debate concerning whether the meaning of external languages is derivative on the meaning of mental states (Grice, 1957) or whether, as behaviourists and others have advocated, the meaning of mental states is derivative on language behaviour (e.g. Quine, 1960). For those who believe the former, as has become orthodox in the foundations of cognitive science (e.g. Fodor, 1975), an account of meaning for internal languages is a necessary prerequisite for providing an account of meaning for external languages too. Thus, elucidating the meaning of internal states may be viewed as *the* fundamental issue in the theory of meaning. The question of what a theory of meaning for internal languages could look like, or whether such a project is feasible at all, has been extensively discussed (Dretske, 1981, 1988; Churchland, 1986; Fodor, 1987, 1990; Stich, 1983; Schiffer, 1987; McGinn, 1989).

Although the classical account of cognition finds the semantics of the external natural language to be individuated by the semantics of the internal language of thought, the argument behind the postulation of the latter goes in the opposite direction, i.e. from external language to internal representation. The argument is based on the observation that natural language is describable in symbolic terms where the symbols correspond to words that can be composed systematically into meaningful complex expressions, sentences, according to a recursively specified syntax.[3] The important link to the internal language of thought is that we use the external language to verbalize (and communicate to others) the contents of our thoughts—or, rather, we use natural language constructs to express the content of our mental representations. However, we can only think (and say) what our mental representations allow us to represent. So, the argument goes, the syntactic as well as semantic systematicity and productivity of the external language must therefore mirror the underlying nature of the internal language of thought by copying its combinatorial syntax and semantics (e.g. *cf.* Fodor & Pylyshyn, 1988). The upshot of this argument with respect to connectionism is that the systematicity of cognitive competences requires mental representations with constituent structure. While

Classical models are defined over structured representations, Fodor & Pylyshyn (1988) and Fodor & McLaughlin (1990) argue that connectionist models *do not* have constituent structure and can therefore not have any compositional semantics. As a result, they conclude that connectionism *ipso facto* does not provide the representational substrate required by a theory of cognition; specifically, it cannot support the systematicity of cognitive capacities. We shall consider this issue further below—but first we turn our attention to the ascription of semantic content to connectionist representations.

3. Connectionism and the Semantics of Primitives

One problem facing theories of meaning relying on the classical account of cognitive processing is that the relation between the primitives of a symbolic system and their semantic content is essentially *arbitrary*. The meaning of the most basic constituents are projected onto them by the observer through semantic interpretation. That is, the meaning of a symbolic system is external to the system itself since it is fundamentally parasitic on the meanings in the head of the observer; or, in more philosophical terms, the atomic symbols have no *intrinsic* meaning. It is therefore always possible to re-interpret the basic symbols, to ascribe them a different content and in this way change the semantic significance of the overall behaviour of the system. The possibility of an externally imposed arbitrary re-interpretation of the representational primitives—originally a cornerstone in Searle's (1980) Chinese room parable (for a discussion, see Harnad, 1989, 1990a; Boden, 1990; Churchland & Churchland, 1990; Dyer, 1990a, b; Searle, 1990; Chalmers, 1992—for a critical review of these, see Christiansen, 1992a), more recently glossed the '*symbol grounding problem*' (Harnad, 1990b)—has plagued the classical paradigm for a long time.

The advent of connectionism has given rise to optimistic expectations regarding a solution to the problem of grounding the semantic primitives of computational systems, be that within a hybrid symbolic/connectionist architecture (e.g. Harnad, 1989, 1990a, b, 1992; Harnad *et al.* 1991) or an entirely connectionist system (e.g. Cottrell 1987; Smolensky, 1988). These expectations have manifested themselves in statements such as, for example, "nets are one possible candidate for the mechanism that learns the sensorimotor invariants that connect arbitrary names (elementary symbols?) to the nonarbitrary shapes of objects" (Harnad *et al.* 1991, p. 1; their brackets), "networks are self-organizing systems that learn to represent the important features of their environment" (Cottrell, 1987, p. 68), and "connectionism offers significant resources for explaining how representations are *about* other phenomena and so possess *intentionality*" (Bechtel, 1989, p. 553). In other words, connectionism allegedly promises a way of providing a computational system with a perceptual 'hook-up' to the external world such that the semantics of its internal representations becomes grounded.

The argument behind this (at least presently) undue optimism with respect to a connectionist grounding of semantic primitives can be expounded as follows (this exposition is a summary of van Gelder's, 1992, discussion). The basic observation is that, through learning, connectionist models (with hidden units) are able to develop internal distributed representations that structurally mirror the structure inherent in the externally given input. More specifically, the vectors that correspond to the individual patterns of activation over the hidden units are often conceived as *points* in a multidimensional state space. The exact location of a given vector is determined by the specific values of its constituents; i.e. by its internal configuration.

As a result, similar vectors are mapped into similar locations in space. The degree of similarity between vectors—the 'distance' between them in space—can be measured using a variety of standard vector comparison methods (e.g. cluster analysis or trajectory analysis). Due to the superpositional and highly distributed nature of the networks in question, representations that are structurally similar—i.e. that have similar internal structure or, more precisely, have similar vector configurations—end up as '*neighbouring*' positions in state space. Thus, structurally related input representations will invoke relatively 'adjacent' representations in hidden unit space.

It is important to notice from a computational perspective that these similarities have *causal significance.* The behaviour of a network, being a complex dynamical system, is causally dependent on the current pattern of activation over the hidden units, i.e. on the current representation's particular location in space. In other words, the specific location in space of a given representation will causally effect how it is processed. Since the internal structure of such distributed representations corresponds systematically and in an essentially non-arbitrary way to the structural configuration of the input representations, allowing us to project any semantic interpretation we might assign the input onto the appropriate positions in vector space, and since variations of position in state space are causally efficacious, the processing of a network can be seen as being determined systematically according to the semantic content of the distributed representations.

Judging from this exposition it would seem to be the case that connectionist representations can be assigned content in an essentially non-arbitrary way, since their internal structure (given successful training) will correlate with structural contingencies in the input and produce a non-arbitrary representation, i.e. connectionist representations appear to be able to possess at least some *bona fide* intrinsic content. However, the internal states of present day connectionist networks appear to be no more 'grounded' than their symbolic counterparts (also *cf.* Bechtel, 1989; Cliff, 1990; Sharkey, 1991). Crucially, the distributed representations in question are only non-arbitrary in relation to the structure of the given input representations, not in relation to what the latter are representations *of,* i.e. the entities they refer to in the outside world. Consequently, similarity is defined as a relation *between* input representations and not as a relation to the appropriate external objects they are to represent. Furthermore, since the input representations provided by the programmer are typically pre-structured and of a highly abstract nature, it is always possible to give a network's input representations a different interpretation, thus changing the projected content of the internal distributed representations. This has been mirrored empirically by the fact that only a few experiments have been carried out with 'real' sensory-type data (in sense of not having been pre-processed by the programmer) and then, as we shall see exemplified below, with a mostly unsuccessful outcome. So, whatever semantic content we might want to ascribe to a particular network, it will always be parasitic on *our* interpretation of that network, i.e. parasitic on the meanings in the head of the observer.

There is, however, a sense in which connectionist representations are non-arbitrary, i.e. the *inter-representational* relations in a network are essentially non-arbitrary. In contrast to symbolic systems in which the atomic symbols have no relation to each other (albeit that complex symbols have non-arbitrary inter-relations), distributed representations are non-arbitrarily related to each other in state space. Whereas atomic symbols designating similar objects have no (non-coincidental) relation to each other, connectionist representations of similar object represen-

tations in the input will end up as neighbouring points in state space. Thus, connectionist networks provide us with a kind of non-arbitrary representational 'shape' that allows a notion of inter-representational similarity. The important ability of connectionist networks to *generalize* derives from these similarity relations between representations corresponding to structurally similar input. Despite the non-arbitrariness of these inter-relations and their grounding of a robust notion of representational similarity, the *extra*-representational links are still fundamentally arbitrary and therefore ungrounded.

3.1. Correlational Semantics

So far, we have pursued the possibility that learnt connectionist representations may be of significance for the theory of meaning as if the meaning of such representations were well understood. In fact, as we shall see, this is not at all the case—meaning in connectionist networks presents a philosophical problem rather than offering philosophical solutions.

Fundamentally, connectionists attach meaning to the states of a network on the basis of what those states *correlate* with. For example, in Hinton's (1986) model of learning family trees, a unit is said to represent nationality or generation in a family if it correlates with these properties in the input. More generally, connectionist units or patterns of activation are viewed as picking out categories, with which they correlate, and which specify their meaning. Thus the network is viewed as acquiring the corresponding concept (of, say, nationality or generation).

For concreteness we shall focus on connectionist models which can be plausibly viewed as involving concept or category learning. In such cases, the learning process can be viewed as learning to correlate the activity of a unit or a pattern of activation over a set of units with some significant aspect of the input (also *cf.*, e.g. Goschke & Koppelberg, 1991; Hatfield, 1991).[4] Examples include unsupervised category learning of all sorts (Carpenter & Grossberg, 1988; Linsker, 1988; Finch & Chater, 1992), and supervised approaches (e.g., Kruschke, 1990) and incidental learning (Hinton, 1986; Elman, 1990, 1991a). This follows tradition in pattern recognition and statistical classification (Duda & Hart, 1973). A similar 'correlational' style of semantics is presupposed within the 'animal concepts' literature (e.g. Herrnstein, Loveland & Cable, 1976; Cerella, 1982; D'Amato & Van Sant, 1988; Chater & Heyes, in submission) and in the interpretation of the activity of real neurons (e.g. Schurg-Pfeiffer & Ewert, 1981). For example, Lea (1984) suggests that to have a concept is to have "... some unique mental structure which is active when and only when an instance of that concept is present in the external, physical environment or when associated concepts are active in the mental environment" (p. 270). According to this view, having the X concept is simply a matter of being able to perceptually discriminate Xs from non-Xs; and such discrimination abilities are just what paradigmatic animal concept experiments aim to test. This correlational account of what it is to have a concept has a counterpart in philosophy as what Jerry Fodor calls the "crude causal theory" of meaning (1987, 1990). A similar correlational account is also closely related to Dretske's (1981) proposal that conceptual structures carry the information that is their content in digital form.

However, there are serious philosophical problems concerning not only a connectionist semantics based on causal correlation but also, in general, the adequacy of correlational semantics as the basis of any theory of meaning. These problems concern the matter of misrepresentation, underdetermination, representing non-ex-

isting things and capturing propositions rather than properties. We will address these problems in turn in the following sections and emphasize their impact on a connectionist semantics.

3.1.1. The problem of error. The fundamental challenge to the correlational view is allowing for the possibility of categorization error. People routinely make both false positive and false negative errors. Mistaking a pattern of shadows for a face at the window is an instance of the former; failing to see a dark figure in the bushes is a case of the latter. Yet the correlational view without some added machinery is unable to countenance the possibility that we have the concept PERSON and that we make such mistakes. For the content of the concept [equally, the meaning of the corresponding state, or for Harnad (1990b, 1992) what the internal state 'names'] is determined by what it correlates with—and the fact of error shows that it does not correlate with instances of people.

As Fodor (1987) points out, the programme of informational semantics within philosophy is concerned with attempting to patch up such problems with correlational accounts. A number of proposals have been made (see, e.g. Stampe, 1977; Millikan, 1984; Dretske, 1986, 1988; Fodor, 1987, 1990; Chater, 1989a; Christiansen, 1992b)—although none are widely considered to be satisfactory. Rather than attempting to survey the range of possible responses, we shall consider just two suggestions about how this problem can be addressed (Fodor, 1984a; Stampe, 1977), which may particularly appeal to connectionists. We contend that other approaches are no more successful.

The first suggestion is that content is fixed during the *learning* of the concept, rather than determined by subsequent performance, outside the learning period, when mistakes may occur (Dretske, 1981). The idea is that the correlation holds within the learning period (fixing the content correctly), but not necessarily afterwards (allowing for the possibility of error). This position is particularly interesting in the present context, in view of the importance of learning in connectionist systems. However, Fodor (1984b) points out that this view is quite unworkable, since all the difficulties for a correlational view recur within the learning period. Let us put aside the difficulty that a single error in the data to be learned (a parent accidentally calling a donkey a horse, perhaps) would leave the learner forever blighted with a non-standard concept, by disrupting the correlation. The real difficulty stems from the fact that the relevant property can never be determined by the training set alone; even if the learner is given perfect feedback about which of a set of things are people and which are not, forming the concept PERSON involves an inductive generalization from a finite set of instances. Which concept has been formed cannot therefore be determined from the correlation observed in the training set alone, since all manner of different properties will fit that training set, but differ elsewhere, such as the pathological PERSON-OR-FACE-LIKE-SHADOWS or PERSON-NOT-IN-CAMOUFLAGE. Which of these concepts has been formed is determined by how the system has *generalized* from the training set, i.e. how it would respond to stimuli outside the training set. Subsequent errors, after the learning period has been completed (assuming that some such boundary can be enforced), demonstrate that generalization has been imperfect; the correlation is violated and the concept has not been learned after all. Thus, appeal to learning fails to reconcile the possibility of learning a concept with proneness to occasional mis-classification.

The second suggestion is that while errors may occur on difficult cases (perhaps when the stimulus is degraded in some way), the correlation that fixes content need

only hold in clear cases. As with appeal to the learning period, the idea is to partition performance into two classes, one in which correlation determines the concept in play (and which is necessarily error-free) and a second class in which the correlation need not be maintained, thus allowing for errors. Unfortunately, however, as Fodor (1990) forcefully argues, what counts as a clear case cannot be specified independent of the concept being learnt. Chater & Heyes (in submission) consider the example of the confusion that commonly occurs at night between a star and the lights of a plane, leading to spurious plane identification. According to appeal to clear cases, it appears legitimate to explain this away, since planes are only confused in this way when they are viewed from a considerable distance and in the dark. In good viewing conditions, perhaps an internal structure does correlate perfectly with the presence of planes. However, while daytime is optimal and nighttime suboptimal for detecting planes, nighttime is optimal for detecting planes or stars (since you can see instances of both at night) and daytime is suboptimal (since only some instances—planes—are visible). So optimality could equally be invoked to argue that the internal mental structure is a PLANE-OR-STAR concept, which correlates properly at night, but imperfectly during the day. The general moral is that the distinction between a class of 'good' cases, where the correlation is supposed to hold, and 'poor' cases, in which error is possible is unconstrained without some independent notion of good and bad case; and such a notion does not appear to be forthcoming.

Thus, according to the correlational position, concepts are defined in such a way that there can be no such thing as 'getting it wrong'. Since the content of the concept is whatever the activity of the mental structure correlates with, misclassification is impossible. The 'learning' and 'optimality' responses are just two of a number of responses which attempt to allow for error by attempting to distinguish two kinds of situations: one kind in which performance determines what content the representation has, and hence what concept it corresponds to; and one kind which is 'non-optimal' (Stampe, 1977; Fodor, 1984a), not 'normal' (in a teleological, non-statistical sense) (Millikan, 1984), or outside the learning period (Dretske, 1981). However, as in the case of optimality, it is extremely difficult to see how to define the distinction between the two classes in a non-circular way.

It may be, of course, that a satisfactory solution to these difficulties can be found—indeed the project of informational semantics is wedded to the hope that it can. While the correlational view, in its least elaborated state, directly ties up with intuitive ascriptions of content to hidden units in connectionist networks, there is, of course, no way of knowing whether a more sophisticated and satisfactory theory of content will tie up in an equally attractive way. Indeed, Fodor's (1987) most recent and ingenious suggestion, which relies on the 'asymmetrical dependence' of counterfactuals underwriting categorization in 'errorless' versus 'error-tolerant' situations, and Millikan's (1984, 1986) advertence to evolutionary considerations do not seem applicable to connectionist networks in any straightforward way.[5]

3.1.2. The problem of underdetermination. In the discussion of error, we noted incidentally that learning a concept from a set of exemplars involves *inductive* inference: inferring a general rule from a set of examples. In neural network terms, this amounts to curve fitting, with the exemplars as the data points and the network architecture specifying the family of curves (e.g. Broomhead & Lowe, 1988; Mackay, 1991). We noted that which inductive rule (i.e. which curve) has been chosen cannot be determined by the training set alone, but is revealed in how the

system would behave given arbitrary test items. It is, of course, notorious that networks trained on a given training set will generalize in unexpected ways. This is particularly true if the network has too many degrees of freedom (i.e. too many weights and biases) relative to the size of the data set and hence does not need to find interesting regularities in the data set (Moody, 1992); networks which show extreme versions of this problem are said to solve their tasks by 'table look-up'.

An early example of the problem of underdetermination with respect to the application of neural networks to 'real', un-processed data is the (now legendary) failure of the optical perceptron.[6] This network was developed in the mid-1960s at the Stanford Research Institute with the purpose of detecting tanks hidden amongst bushes. It consisted of a large number of optical masks with photodetectors that produced weighted sums of photographic input. The network was trained successfully to differentiate between photos of bushes and photos with tanks amongst the bushes. It was also able to generalize to photos that had not been presented to it previously during the learning phase. To be certain that the network had really learned to recognize tanks, a new set of photos was taken and presented to the network. However, this time the network failed completely to categorize the photos. Apparently, the network had not learned to recognize tanks but to differentiate between photos of different light intensity. A closer examination of the photos used to train the network indicated that there was a large difference in intensity between the batch of photos which had tanks in them and the batch of photos with bushes only.

Yet the problem of underdetermination is deeper than these examples suggest—it cannot be resolved by generalization tests, however ingenious. Consider, for example, a network successfully trained on the above 'tank discrimination' task. Suppose that we discover exactly which complex structural properties of the photos the network has learned to respond to. The network might respond positively when presented with any stimulus containing one of a set of contour relationships, hues and so on. Let us assume that stimuli which have these properties usually look, to the human eye, like a tank. Indeed, we may label, after Fodor (1990)—who calls this the '*disjunction problem*'—the relevant complex constellation of properties of the stimulus 'that-tanky-look'. It is likely that such findings would prompt the announcement that we now know the perceptual basis upon which this kind of network distinguishes tanks from non-tanks, categorizes tanks or applies the concept TANK.

This portrayal of the data certainly seems to be legitimate, but unfortunately, there appear to be equally legitimate alternatives. On the one hand, it might be argued that such research really shows that the network does not have the concept TANK at all, but merely the concept of THAT-TANKY-LOOK. Indeed, the latter interpretation might hypothetically be supported by data indicating that the net can be foxed by camouflaged tanks or theatre mock-up tanks. So, although in everyday life tanks and instances of 'that-tanky-look' are perfectly correlated, it is clear from the cases in which they are not correlated that it is the latter, rather than the former, to which the net is responding. The *prima facie* viability of this option then presents a dilemma. Either sensitivity to mere correlates of this property (such as 'that-tanky-look'), rather than to the property itself, is sufficient for the possession of the corresponding concept (TANK) or it is not.

If sensitivity to mere correlates of the relevant property is not sufficient, then there is both good news and bad news. The good news is that an investigation of the bases of network discrimination (filling out what 'that-tanky-look' amounts to)

automatically specifies what concept or category the network is using. The bad news is that the network must be ascribed a concept the referent of which is ultimately a state of its input 'sensors' (e.g. THAT-AN-INPUT-WITH-THAT-TANKY-LOOK-IS-PROJECTED-ON-THE-SENSORS). In the light of these sceptical considerations, it seems that however the data turned out, networks could only be expected to learn concepts of THAT-TANKY-LOOK/BUSH-LIKE/TREEISH variety, rather than *bona fide* concepts.

Suppose, on the other hand, that one assumes that in order to possess a concept, it is sufficient to be sensitive to perceptual correlates of the corresponding property. That is, to have a concept of TANK, it is necessary only to be sensitive to some correlated complex perceptual property. We will continue to call any such hypothetical property 'that-tanky-look'. Again, there is both good news and bad news. The good news is that this more lenient view allows the possibility of networks learning everyday concepts, since this requires detecting, say, tanks reliably most of the time. So what is the bad news? While we can ascribe concepts of the kind TANK, BUSH or TREE to a network, we can only do so relative to some characterization of the environment. Consequently, very different concepts may be ascribed depending on which characterization of the environment is chosen. Since, according to this view, concept ascription is a matter of correlation and correlation is fundamentally relative to a specification of the domains of values being correlated, concept ascription must also be domain-relative.

This may be illustrated by visuo-motor coordination in frog and toad. Much is known about the frog's visual system, the frog's range of motor outputs and how the two are related (Lettvin *et al.*, 1959; Ingle, 1983; Schurg-Pfeiffer & Ewert, 1981). As an idealization, let us assume that we know precisely which visual stimuli will elicit the predatory movement 'snapping'. In particular, suppose that it is triggered by the projection of any dark, round, moving blobs within a certain range of sizes, projected onto the frog's retina. So, if the frog were sitting in a Scottish stream and the projection of moving black blobs correlated with the presence or passage of flies across the stream, then, according to this 'lenient' approach to concept ascription, the frog might legitimately be described as having the concept FLY. However, in this situation, the frog may also be ascribed more specific or more general concepts. After all, since, by hypothesis, the only passing flies will be natives of Scotland, the blobs would correlate just as well with 'Scottish flies' as with 'flies'. Similarly, again by hypothesis, the only passing flies would also be the only passing flying insects and hence the frog might be described as having the concept FLYING-INSECT. The range of possible concept ascriptions can be extended at will and hence appear to be completely unconstrained (see Chater & Heyes, in submission, for discussion of these issues in the context of animal concepts).

3.1.3. The problem of non-existing entities. A further problem for causal/correlational accounts of meaning is explaining the origin of the meaning of terms such as UNICORN or EPICYCLE which have no instances and hence cannot be either causally implicated in producing, or correlated with, internal states (see, e.g. Fodor, 1987). These symbols cannot be 'grounded' by some state of a network which comes to correlate the presence of unicorns or epicycles in the environment; for these are *never* present in the environment—they do not exist. Problems with non-existent universals has plagued philosophy since Hume. The only proposed solution for a causal/correlational view is that the meaning of non-existents is composed out of the meaning of more primitive terms, which do exist. So, the story goes, *unicorn* means

horse with a central horn, and since horses, central things and horns all exist, then unicorn inherits its meaning from them.

This view presupposes that terms for things which do not exist can be defined in terms of things that do. This position appears to have the rather radical consequence that *every* term must have a definition. For suppose that a term X does not have a definition; then it must refer to something real; hence Xs must exist. Thus a semantic fact (concerning definability) appears to be revealing about a metaphysical fact (whether there are Xs). On the face of it, this means that we could learn what there is in the world, simply by examining language, which seems to be absurd [although arguments from semantics to metaphysics have been attempted (e.g. Kripke, 1972) and rebutted (Salmon, 1982)]. So it seems that we must conclude that *every* term must be definable in terms of other terms. The thesis that some terms have good definitions is highly controversial; the thesis that *all* terms do is so radical that it has not, to our knowledge, ever been advanced.

3.1.4. Propositions and properties. Whether or not it is possible to patch up the informational view to get around the preceding difficulties, the correlational account, *construed as a method of fixing the meaning of concepts* is in any case victim to a much more fundamental problem (Chater, 1989a).[7] The problem is that while the correlational approach at least promises to provide an account of how internal states (e.g. internal states of a network) can represent propositions, it provides no account at all concerning how they can represent *properties.* Since concepts are mental representations which stand for properties, this also means that the correlational view provides no account of what it is for an internal state to correspond to a concept.

So far, we have been relying on the intuition that a state will come to represent the property of, say, being red, if that state is active in the presence of redness and not otherwise. Speaking roughly, the state is supposed to correlate with the property of being red. However, as stated, this is simply incoherent—how can a *state,* which is located in space and time, correlate with a *property,* which is an abstract object, independent of space and time? The answer appears to be obvious—the state of the system correlates not with the property itself, but with *tokens* of that property.

This, however, is not good enough. Sensitivity to tokens of properties presupposes the ability to recognize the relevant *tokens.* What it is to recognize a token and that a token is an instance of a property is a matter concerning which the correlational theory is silent. Let us elucidate this using the example of detecting redness. The red-detecting unit responds each time red is in view, i.e. the state of the cell correlates with a state of the world, that red is present (and in the visual field, and not too distant to be seen, and not occluded by another object—let us put these complications to one side). However, is the cell sensitive to tokens of redness (e.g. that this pen or that cup is red, but that that jumper is not)? It is not—it does not signify that any particular token is red. If anything the state of the detector has existential force: it represents the fact that some token or other is red. It certainly does not ascribe the property of redness to any particular token.

This is not just a quibble—the difference between being able to represent the unanalyzed (in philosophical terminology *holophrastic*) proposition that red is present, and being able to segregate parts of the world and selectively ascribe properties to them is enormous. The representations licensed by the correlational view amount to a set of binary features, which specify red/not-red, fly/not-fly, person/not-person, with, of course, all the problems notoriously inherent in such a primitive representation. There is, for example, no way, even in principle, of binding

these features together (there is no way of representing that it is the person who has a red face and that the fly sits on the end of her nose), representing which tokens share particular properties, how many red things there are in a particular scene and so on. To achieve this, we require a *language* in which to couch *structured descriptions* of the world, segregating the world into a complex set of tokens, each of which can individually be ascribed properties. The advantages of a structured description over simple binary feature representations are too well known in the literature on computational approaches to perception and other areas of cognition to bear repetition (see, e.g. Marr, 1982). The important point is that it is only for systems with such structured representations that we can talk of properties, rather than whole, unanalyzed propositions, being represented at all and that correlational accounts are equipped only to fix the meaning of whole propositions. In particular, then, the correlational account (and *mutatis mutandis* causal accounts of related sorts) cannot fix the meaning of primitives of an internal system of (structured) representation. So, for example, the hope that a neural network could effect so-called *symbol grounding* by learning appropriate correlations between states of the network and aspects of the world appears to be illusory.

The upshot of this discussion should not, perhaps, be surprising. Unless a system embodies a structured internal language with an associated set of primitives, it is difficult to see how it could possibly throw light on the semantics of those primitives. The networks that we have considered so far are not concerned with structured representation. In short, no internal language and no symbol system, no symbols to ground. The conclusion is, then, not that network computation is necessarily irrelevant to the theory of meaning for semantic primitives; rather, it is that connectionism can be relevant only if structured representations are somehow embodied in a network. It is therefore to this issue that we now turn.

4. Learning Complex Representations in Connectionist Systems

One way of approaching the problem of dealing with structured representation in connectionist models is to 'hardwire' symbolic structures directly into the architecture or the network. Much early work in, for example, connectionist knowledge representation (e.g. Hinton, 1981; Touretzky & Hinton, 1985; Rumelhart *et al.*, 1986; Derthick, 1987) and natural language processing (e.g. McClelland & Kawamoto, 1986) adopted this implementational approach. Although such connectionist re-implementations of symbolic systems might have interesting computational properties and even be illuminating regarding the appropriateness of a particular style of symbolic model for distributed computation (Chater & Oaksford, 1990a), they do not appear to have much philosophical significance (if any). However, there is the promise that connectionism may be able to do more than simply implement symbolic representations and processes; in particular, that networks may be able to *learn* to form and use structured representations. The most interesting models of this sort typically focus on learning quite constrained aspects of natural language syntax. These models can be divided into two classes, depending on whether preprocessed sentence structures or simply bare sentences are presented.

The less radical class (e.g. Hanson & Kegl, 1987; Pollack, 1988, 1990; Stolcke, 1991; Sopena 1991) presupposes that the syntactic structure of each sentence to be learnt is given. The task of the network is to find the grammar which fits these example structures. This means that the structured aspects of language are not themselves learned by observation, but are built in. These models are related to

statistical models approaches to language learning such as stochastic context free grammars (Brill *et al.*, 1990; Jelinek *et al.*, 1990) in which learning sets the probabilities of each grammar rule in a prespecified context-free grammar, from a corpus of parsed sentences.

The more radical models have taken on a much harder task, that of learning syntactic structure from strings of words, with no prior assumption of a particular syntactic structure to the grammar. The most influential approach is to train simple recurrent networks (SRNs) developed by Jordan (1986) and Elman (1988). These networks provide a powerful tool with which to model the learning of many aspects of linguistic structure (e.g. Elman, 1990, 1991a; Norris, 1990; Cottrell & Plunkett, 1991; Shillcock *et al.*, 1991); there has also been some exploration of their computational properties (Chater, 1989b; Cleeremans *et al.*, 1989; Servan-Schreiber *et al.*, 1989, 1991; Chater & Conkey, 1992). The presence of recurrent connections allows past activation to influence current output, which means that output can respond to sequential structure in the input. The extent to which such networks can be taught to learn interesting sequential structure depends on the learning algorithm employed. A natural approach is to apply the backpropagation training algorithm which has proved so successful in training non-recurrent feedforward networks to learn interesting static input–output patterns.

It is fair to say that these radical models have so far reached only a modest level of performance. In general, it seems to be possible to learn simple finite state grammars, but more complex grammars, such as phrase structure grammars have not been learnt [although Elman (1991a) claims to be able to train a SRN to learn a limited instance of recursion]. The gulf between finite state and phrase structure grammars is a vast one—and it is not clear whether current network models will be able to cross it. It may be that only by pursuing the less radical line, by building in more structure into the network itself, that complex linguistic structures will be learnable. Given the negative results of standard language learning theory (e.g. Gold, 1967; Pinker, 1979, 1984; Osherson *et al.*, 1986), which show that even finite state language cannot be reliably learned from (positive) examples alone, there is reason for scepticism regarding the possibility of a connectionist breakthrough [although see Elman (1991b) for the opposite view]. It is, however, simply too early to tell.

Having pondered the difficulty of connectionist modelling of structured representation in natural language, we may suspect that connectionism leaves the general issues of structured representation and the associated compositional semantics unresolved. As we shall see now, there are indeed indications that this might be the case.

4.1. Connectionism and Compositionality

We noted above that the allegedly most revolutionary consequences of connectionism concerns the nature of connectionist representation (also *cf.*, e.g. Bechtel, 1989; Haugeland, 1991; Sharkey, 1991, 1992; van Gelder, 1991; Niklasson & Sharkey, 1992—but see Hanson & Burr, 1990; Cummins, 1991 for different views). Typically the focus has been on devising connectionist networks which are able to deal with problems for which the symbolic approach invokes syntactically structured representations. This is clearly exhibited in the debate initiated by Fodor's & Pylyshyn's (1988) attack on connectionism (see, e.g. Smolensky 1987, 1988; Chalmers, 1990b; Chater & Oaksford 1990a; Fodor & McLaughlin, 1990;

Oaksford, Chater & Stenning, 1990; van Gelder, 1990, 1992). In contrast to the intensive studies of structural combination of constituents in connectionist models not much have been said about *semantic composition* in connectionist networks—unless in rather vague terms (e.g. Goschke & Koppelberg, 1991). Of course, as we noted above, without some kind of semantic interpretation, we cannot view a system, whether symbolic or connectionist, as *processing information* or *computing* at all; in the present context, this might involve a compositional semantics for the structured representation, to show how the meaning of complex structures is related to the meaning of their parts. We shall briefly return to the question of semantic interpretation below.

It has been suggested that the classical notion of compositionality may be unnecessarily restrictive from the point of view of connectionist systems (i.e. the classical understanding of compositionality may induce an instance of the incommensurability trap—forcing connectionist systems into an inappropriate framework). This classical notion is labelled as *concatenative* (or 'syntactic') compositionality, which "must preserve tokens of an expression's constituents (and the sequential relations among tokens) in the expression itself" (van Gelder, 1990, p. 360).

A broader notion, *functional* compositionality, does not demand the preservation of constituents in compound expressions. What *is* needed is a general and reliable mechanism that can produce composite expressions from arbitrary constituents and later decompose them back into their original constituents. As an example of functional compositionality, van Gelder (1990) points to Gödel numbering, which is a one-to-one correspondence between logical formulae and the natural numbers. For instance, on a given scheme the proposition P will be assigned the Gödel number 32, whereas a logical expression involving P as a constituent, say (P&Q), would be assigned the Gödel number 51342984000. It is clear that the Gödel number for (P&Q) does not directly (or syntactically) contain the Gödel number for P. Still, by applying the prime decomposition theorem we can easily determine the Gödel numbers for its primitive constituents. Thus, we have constituency relations without concatenative compositionality. Since distributed networks using superimposed representation effectively 'destroy' the constituents of composite input tokens, they do not qualify as having concatenative compositionality. However, this is not irreversible because the original constituents can be recreated in the output.[8]

There is a danger that this would leave connectionist representations with the same status as, for example, data-compressed, or otherwise encrypted, files on a standard computer—as being useful only as storage but not for processing. For a genuinely connectionist account of representing and processing with structured representations, it is necessary to be able to manipulate the functionally compositional representations *directly* as van Gelder stresses. In the case of Gödel numbering, operations which are sensitive to compositional structure (e.g. inferences) will not correspond to a (readily specifiable) function at the arithmetic level. Hence, performing logical inference over Gödel numbers is a rather hopeless endeavour. Notice too, the compositional *semantics* which can be easily defined over logical representations will have no (readily specifiable) analog at the level of Gödel numbers.

What is important, from the present point of view, is whether or not connectionist networks can handle (and, in particular, learn to handle) problems which are standardly viewed as requiring structured representations. That is, can connectionist representations attain what we shall call '*apparent*' compositionality. If apparent

compositionality can be learnt, then there are two possibilities concerning the nature of the representations that the network employs. It could be that, on close analysis, the net is found to have devised a standard, concatenative compositional representation. Alternatively, the network might behave *as if* it used structured representations, without using structured representations at all. In the former case, it would seem appropriate to say that the network representations are compositional (in the standard sense); in the latter, that the network is not using a compositional representation (also in the standard sense). What is required, it appears, is not a new notion of compositionality, but the attempt to devise networks which can behave as if they had structured representations, followed by an analysis of their workings. Of course, there is a third possibility: that representations within networks do implement compositionality, but in some heretofore unknown way, unlike that used by classical systems (with appropriate operations over it, and an appropriate semantics). *This* possibility would cause us to revise the notion of compositionality, much as the discovery of non-Euclidean geometry enlarged and changed the notion of straight lines, parallel and so on. It will only be possible to develop a specifically connectionist notion of compositionality, or even know if this possibility is coherent at all, *post hoc*, i.e. by analyzing networks that exhibit apparent compositionality.[9] In other words, what kind of compositionality we should ascribe connectionist representations is an empirical question, which can only be answered by empirical investigation.

Recently, research efforts have therefore been made towards defining operations that work directly on the encoded distributed representations themselves, instead of their decomposed constituents. Chalmers (1990a) devised a method by which a simple feed-forward, backpropagation network—dubbed a *transformation* network (TN)—was to manipulate compact distributed representations of active and passive sentences according to their syntactical structure. First, a recursive auto-associative memory (RAAM) network (Pollack, 1988) was trained to encode distributed representations of the sentence structures. Chalmers then trained the TN to transform compact representations of active sentences into compact representations of passive sentences, i.e. he trained the network to associate the RAAM-encoded distributed representations of active sentences with their distributed passive counterpart. In a similar vein, Niklasson & Sharkey (1992) successfully applied the same combination of RAAM[10] and TN to (a subpart of) the domain of logical axioms. These empirical investigations have shown that it is possible to devise models, such as the TN, that can manipulate the compact distributed representations in a structure sensitive way. However, with respect to the semantics of these encoded representations, we still have to decompose than into their symbolic parts *before* we can perform any semantic interpretation of them. What connectionism is in need of is some kind of compositional semantics devised at the level of the compact distributed representations and the operations defined directly over them; that is, a *bona fide* connectionist semantics that does not have to revert to semantic interpretation of the *decoded* constituents on the symbolic level.

In closing this section, it is worth mentioning that when addressing the issue of connectionist compositionality there is a potential danger of falling into the incommensurability trap. As pointed out by Sharkey (1991), the division between semantics and structural considerations might be somewhat artificial, since such a division seems to be collapsed in much connectionist research. The situation could be seen to parallel that of the classical/connectionist debate concerning implicit vs. explicit rules. When a connectionist model behaves *as if* it has rules, although no

rules have been programmed into it, does that warrant saying that the model has 'implicit' (or 'fuzzy') rules? Such talk about implicit rules is in danger of forcing connectionism into a symbolic mold by trying to apply a particular concept, i.e. the classical notion of a computationally efficacious rule, to connectionism. On this view, even our own notion of apparent compositionality could get us trapped in the claws of incommensurability. Nevertheless, bearing this in mind, re-interpretation of old terminology seems to be the only productive way forward for a research programme still in its infancy. Furthermore, whereas the ability of nets to deal with structured representations is equivocal, their aptness for learning seems to be more clearcut; so perhaps, as we shall see next, it is with respect to questions of learning and nativism that their principal philosophical significance resides.

5. Learning and Nativism

Symbolic models of cognition appear to offer relatively little scope for learning. For example, for Fodor & Pylyshyn (1988) cognition is mechanized proof theory over a data-base of facts couched in a language of thought. Such a system can learn by adding facts to its data-base, but it is not clear how to learn a more elaborate internal language in which facts can be expressed. Thus, according to the symbolic view of the mind, the set of possible mental states and concepts appears to be fixed. This line of radical nativism has been pressed by Fodor (1975, 1981; and the various contributions to Piatelli-Palmarini, 1980), who argues that concept *learning* is, in a certain sense, impossible and that a nativist conclusion with respect to the language of thought is therefore inevitable. The argument is simply that learning is a matter of generating and testing hypotheses and hence that any hypothesis that can be framed must already be representable by the system. Consider, for example, a child learning the meaning of the word 'dog'. To be able to generate the correct hypothesis at all, the child must be able to internally represent some predicate which means *dog*. However, the argument goes, this requires that the child already has the concept DOG. This leads to the conclusion that learning the meaning of a new word does not involve *concept* learning at all, but simply involves learning to associate internal and external languages appropriately. The expressive power of the internal language is fixed; and thus must be specified innately.

Do the nativist arguments apply equally to connectionist models as well? Certainly connectionist learning can be viewed as a kind of hypothesis generation and test—the hypotheses are embedded in the weights of the network, the test is the measure of network performance (such as sum-squared error), and the procedure for generating new hypotheses, given the successes or failures of past hypotheses, is given by the learning algorithm. Hence, the above argument applies, just as before: anything that a network can represent after learning, must have been generated as a hypothesis; hence it must have been possible to represent it prior to learning; hence the representational power of a connectionist system cannot change through learning. An obvious objection is that the representation genuinely has been learnt during training and was not present to start with—after all, the initial weight values are typically set randomly. While this is correct, it does not contradict Fodor's argument, which concerns not what a network happens to represent, but what it is *able* to represent. Fodor is arguing that any hypothesis and test procedure cannot increase what is *potentially* representable. For example, a simple perceptron (i.e. a network with only one layer of adjustable weights and a single output unit which is on if the input exceeds a certain threshold and off otherwise) is able to represent only

linearly separable categories (Minsky & Papert, 1969). This is a limitation of the architecture—it specifies what the network is able to learn in principle and cannot be altered by learning. According to this line of thinking, neural network and symbolic systems are both equally trapped at a fixed level of representational power, which cannot be increased by learning. The problems that Fodor raises appear to apply equally to both cases.

There is a sense in which Fodor's argument, when considered at the most general level, becomes entirely trivial. Fodor notes that what can, in principle, be represented by a system cannot increase. However, in the same uninteresting sense, potential of *any* sort cannot increase. In particular, for a system to learn or do anything, it must necessarily have had the potential to learn or do that thing in the first place, i.e. a system can only fulfill its potential, not exceed it. For example, if Fred has the potential to jump 6 feet high, then he must have had that potential (in the relevant, vacuous sense) when he was a child; or, to give another example, learning geography or physics does not change the potential for learning geography or physics—for anything that is actually learned must have been potentially learnable beforehand. It is in this trivial sense that the representational potential of a network or symbolic system is fixed (strictly, representational potential, like all potential need not be static, but can only decrease). In addition, Fodor's argument, couched in general terms so as to apply to any learning system (specifically any system which learns by hypothesis generation and test, although it is not at all clear that there are learning methods which do not conform to this stricture) is thus analogous to an even more general argument that learning *anything* is in a certain sense impossible.

The real issue, then, is not whether representational potential can increase, but how a system can learn to represent new things. That is, we want to be able to say that a system which has learnt to distinguish Xs from non-Xs after a long period of training, is now able to represent a distinction that it could not previously represent. This intuition applies equally well to symbolic and connectionist systems. So, for example, Winston's (1975) classic model of learning the structure of arches from instances and non-instances involves a system composing representational primitives in a new way (see also Lenat, 1982). Or in a connectionist context, a system which learns to divide words into syntactically interesting categories from raw data (Elman, 1990; Finch & Chater, 1992) involves complex weight adjustment to represent these distinctions. Symbolic models, by presupposing an entire system of representation, appear to involve stronger nativist assumptions that connectionist networks (which only presuppose a particular network architecture and a choice of input and output representation). However, as Fodor (1981) points out, the fact that there appear to be no or almost no good definitions of terms (of non-mathematical domains, at least) means that the advocate of a symbolic approach to learning is forced into a more nativist position still: if a term has no good definition, it cannot be constructed by composing a set of primitives according to the methods of symbolic learning, and hence it must be innate. Thus, while a trivial argument for representational nativism applies to symbolic and connectionist systems alike, the real nativist considerations apply only to the former. Given the flexibility of human cognition and development, such nativism is extremely difficult to accept. In learning music, physics or sailing we seem to learn entirely new sets of concepts which are at once not definable in terms of previous understanding and which it seems highly implausible to view as innate.

Together, the discussion of learning in this section and the discussion of structured representation in the previous section leads to the conclusion that symbolic

models are good at structured representation and poor at learning; whereas connectionist networks are good at learning but (at least presently) poor at structured representation. This suggests that the connectionist project of attempting to learn structured representations may provide a bridge between the two approaches—providing an account of how networks can embody structured representations, and providing an account of how the representational power of a symbolic system can genuinely increase (in the most optimistic scenario, increase from nothing). Whether or not it will be possible to, in this way, vindicate the representational power of neural nets and dissolve the symbolic theorists' enforced nativism cannot be known *a priori*. As we saw in the last section, early signs are at best equivocal. However, what is clear is that the challenge of learning complex structured representations is fundamental to elucidating the philosophical implications of connectionism.

6. Conclusion

We began this paper by noting the *prima facie* anomaly of the philosophical excitement surrounding connectionism, particularly in regard to the theory of meaning. Instead, we have put forward what we believe to be a more realistic characterization of the present stage of the connectionist research programme, arguing that much of this excitement is indeed unfounded. In the light of our comments, a natural reaction might be to suggest that the theory of meaning should ignore connectionism. As we have seen, connectionism has so far not solved the problem of how primitive representations can be grounded—in fact, the interpretation of states of connectionist networks has usually presupposed a familiar correlational approach to fixing the content of semantic primitives, a view with a range of serious difficulties. Equally, networks have not yet given a fresh perspective on how complex representations can be built out of simpler components—either traditional compositional mechanisms are built into a connectionist system, or, if learning is used, the resulting system rarely appears to be a genuine substitute for the symbolic alternative.

Nonetheless, we suspect that it would be a mistake for the theory of meaning to neglect future connectionist developments. Connectionism is still in its infancy and the representations that can be developed may become increasingly philosophically interesting, particularly with regard to connectionist models of tasks usually viewed as involving structured symbolic representations. In this connection, it is important to notice that there is a growing bulk of evidence from research into concepts and categorization which argues against the straightforward mode of semantic concatenative compositionality of the symbolic approach (e.g. Murphy & Medin, 1985; Barsalou, 1987; Brooks, 1987; Lakoff, 1987; McCauley, 1987; Medin & Wattenmaker, 1987; Hampton, 1988; Medin & Shoben, 1988; Chater *et al.*, 1990—for an overview, see Christiansen, 1992d). An increasing number of connectionists (e.g. Hofstadter, 1985; Bechtel, 1989; Goschke & Koppelberg, 1991) argue that connectionism might be able to accommodate these results by devising a different, essentially non-classical, way of composing complex meanings from more primitive parts.

Still, although it may be a mistake to expect connectionism to *solve philosophical problems*, it may pose important *philosophical challenges*. For example, providing a representational account of the operation of a connectionist system which has *self-organized* into a system using complex, structured representations, and the way in which it developed, would provide an extremely interesting and important test-bed

for accounts of meaning. However, it should be noticed that there are problems with the theory of meaning *per se*—not only with respect to classical and connectionist models—and it could be the case that no such theory is possible at all (*cf.* Schiffer, 1987). In any case, connectionism might, at least, provide a way out of the representational nativism into which classical symbolic theorists are forced and perhaps open the way for a more acceptable account of cognitive development. Of course, which philosophical challenges connectionism will generate, as well as its potential significance as a new metaphor for the mind, cannot be decided *a priori* through philosophical investigation. Rather, it is an empirical issue—only time, and the vigorous development of connectionist research techniques, will tell.

Acknowledgements

M.H.C. is supported by award no. V910048 from the Danish Research Academy. N.C. is partially supported by grant no. MRC FPG 9024590 from the Joint Councils Initiative in Cognitive Science/HCI.

Notes

1. Since only fully distributed, superpositional networks, trained through some kind of learning procedure, are fundamentally different from the classical symbolic models (e.g. *cf.* Sharkey, 1991; van Gelder, 1991, 1992), we will only address the potential philosophical implications of this kind of connectionist models.
2. For example, much of the criticism of connectionism launched by Fodor & McLaughlin (1990) as well as Fodor & Pylyshyn (1988) does not stem from inconsistencies or incoherence *within* the theoretical framework of connectionism. Instead, it stems from the tendency on behalf of Fodor and collaborators to couch connectionism in the terminology of the classical processing paradigm (also *cf.* van Gelder, 1992). Similarly, another non-classical approach to cognition—situation theory—has also been victim of the same kind of terminologically based criticism: "Fodor thinks that computation is formal. So when I argue that thought is not *formal*, he annoyingly charges me with claiming that thought are not *computational*. I suppose Fodor is so caught up in his own identification of formal with computational as to be unable to maintain the distinction" (Barwise, 1989, pp. 156–157).
3. For a connectionist-inspired criticism of the alleged necessity of recursion in accounts of natural language behaviour, see Christiansen (1992c).
4. Although we recognize the importance of *superposition* in connectionist models (e.g. van Gelder, 1991, 1992; Sharkey, 1992), this particular issue is orthogonal to the following discussions of the philosophical problems facing a connectionist semantics based on correlational content. Superposition is essentially about *inter-representational* relations, not about the relationship between representations and the external world.
5. However, connections between networks and evolution (Bechtel, 1989; Goschke & Koppelberg, 1991) or biological function (Hatfield, 1991) may be relevant here.
6. Thanks to Marvin Minsky (personal communication) for providing the details of this piece of early neural network research.
7. Thanks are due to Jerry Seligman for extensive discussion of this point.
8. The idea is that representations can have *functional compositionality* in virtue of standing in an appropriate one-to-one correspondence with representations which have concatenative compositionality. The general form of this usage is: given any two sets X (say, the set of logical formulae) and Y (say, the Gödel numbers for these formulae) which stand in one-to-one correspondence, any property P (say, being compositional) of X could be said to licence Y's having the property *functional* P (say, being functionally compositional). That is, given any two sets in one-to-one correspondence, the properties of one will be the 'functional' properties of the other and *vice versa*. So, for example, given a one-to-one mapping between the set of even numbers and the set of odd numbers, the latter could be said to be *functionally divisible by 2*.
9. Of course, it is likely that any such notion would be included as a subclass of functional compositionality (as is the case with concatenative compositionality)—but functional compositionality *per se* does not put us any further forward to finding such a notion.

10. Actually, they applied a slightly modified version of the RAAM which in addition to the encoding and decoding of distributed representations also was trained to distinguish whether the input–output representations were atomic (i.e. not distributed) or complex (i.e. distributed).

References

Barsalou, L. W. (1987) The instability of graded structure: implications for the nature of concepts. In U. Neisser (Ed.), *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization.* Cambridge: Cambridge University Press.

Barwise, J. (1989) Unburdening the Language of Thought. In *The Situation in Logic.* Stanford, CA: Centre for the Study of Language and Information.

Bechtel, W. (1989) Connectionism and Intentionality. In *Proceedings of the 11th Annual Meeting of the Cognitive Science Society.* Hillsdale, NJ: Lawrence Erlbaum.

Bechtel, W. & Abrahamsen, A. (1991) *Connectionism and the Mind: An Introduction to Parallel Distributed Processing in Networks.* Oxford: Basil Blackwell.

Boden, M. (1990) Escaping from the Chinese room. In M. Boden (Ed.), *The Philosophy of Artificial Intelligence.* Oxford: Oxford University Press.

Brill, E., Magerman, D., Marcus, M. & Santorini, B. (1990) Deducing linguistic structure from the statistics of large corpora. In *DARPA Speech and Natural Language Workshop.* Hidden Valley, PA: Morgan Kaufmann.

Broomhead, D. S. & Lowe, D. (1988) Radial basis functions, multi-variable functional interpolation and adaptive networks. RSRE Memorandum 4148. Malvern, Royal Signals Research Establishment.

Brooks, L. R. (1987) Decentralized control of categorization: the role of prior processing episodes. In U. Neisser (Ed.), *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization.* Cambridge: Cambridge University Press.

Carpenter, G. & Grossberg, S. (1988) The ART of adaptive pattern recognition by a self-organising neural network. *Computer,* **21** (3), 77–90.

Cerella, J. (1982) Mechanisms of concept formation in the pigeon. In D. J. Ingle, M. A. Goodale & R. J. W. Mansfield (Eds), *Analysis of Visual Behaviour.* Cambridge, MA: MIT Press, pp. 241–262.

Chalmers, D. J. (1990a) Syntactic transformations on distributed representations. *Connection Science,* **2,** 53–62.

Chalmers, D. J. (1990b) Why Fodor and Pylyshyn were wrong: the simplest refutation. In *Proceedings of the 12th Annual Meeting of the Cognitive Science Society.* Hillsdale, NJ: Lawrence Erlbaum, pp. 340–347.

Chalmers, D. J. (1992) Subsymbolic computation and the Chinese room. In J. Dinsmore (Ed.), *Closing the Gap: The Symbolic and Connectionist Paradigms in Cognitive Science.* Hillsdale, NJ: Lawrence Erlbaum.

Chater, N. (1989a) *Information and information processing.* PhD Thesis, Centre for Cognitive Science, University of Edinburgh.

Chater, N. (1989b) Learning to respond to structures in time. Technical report RIPRREP/1000/62/89. Malvern: Research Initiative in Pattern Recognition.

Chater, N. & Conkey, P. (1992) Finding linguistic structure with recurrent neural networks. In *Proceedings of the 14th Annual Meeting of the Cognitive Science Society,* Indiana University, Bloomington, July/August, pp. 402–407.

Chater, N. & Heyes, C. Animal Concepts: Content and Discontent, in press.

Chater, N., Lyon, K. & Myers, T. (1990) Why are conjunctive categories overextended? *Journal of Experimental Psychology: Learning of Experimental Psychology: Learning, Memory and Cognition,* **16,** 497–508.

Chater, N. & Oaksford, M. R. (1990a) Autonomy, implementation and cognitive architecture: a reply to Fodor and Pylyshyn. *Cognition,* **34,** 93–107.

Chater, N. & Oaksford, M. (1990b) Logicist cognitive science and the falsity of commen-sense theories. Technical report UWB-CNU-TR-90-4. Bangor: Cognitive Neurocomputing Unit, University of Wales.

Christiansen, M. (1992a) Intentionality and representation: a view from the Chinese room. In N. Chater & M. Pickering (Eds), *Representation and Folk Psychology.* Edinburgh Working Papers in Cognitive Science, Centre for Cognitive Science, University of Edinburgh.

Christiansen, M. (1992b) A new brand of folk psychology. In N. Chater & M. Pickering (Eds), *Representation and Folk Psychology.* Edinburgh Working Papers in Cognitive Science, Centre for Cognitive Science, University of Edinburgh.

Christiansen, M. (1992c) The (non)necessity of recursion in natural language processing. In *Proceedings of the 14th Annual Meeting of the Cognitive Science Society,* Indiana University, Bloomington, July/August, pp. 665–670.

Christiansen, M. (1992d) Beyond localist concept representation. Unpublished manuscript: Centre for Cognitive Science, University of Edinburgh.

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

Churchland, P. M. & Churchland, P. S. (1990) Could a machine think? *Scientific American,* **262,** 26–31.

Churchland, P. S. (1986) *Neurophilosophy: Toward a Unified Understanding of the Mind-Brain.* Cambridge, MA: MIT Press.

Clark, A. (1989) *Microcognition: Philosophy, Cognitive Science and Parallel Distributed Processing.* Cambridge, MA: MIT Press.

Cleeremans, A., Servan-Schreiber, D. & McClelland, J. L. (1989) Finite state automata and simple recurrent networks. *Neural Computation,* **1,** 372–381.

Cliff, D. T. (1990) Computational neuroethology: a provisional manifesto. Technical report CSRP-162. Brighton: School of Cognitive and Computing Sciences, University of Sussex.

Cottrell, G. W. (1987) Toward a connectionist semantics. *Theoretical Issues in Natural Language Processing 3.* New Mexico: Association for Computational Linguistics, University of New Mexico.

Cottrell, G. W. & Plunkett, K. (1991) Learning the past tense in a recurrent network: acquiring the mapping from meanings to sounds. In *Proceedings of the 13th Annual Meeting of the Cognitive Science Society.* Hillsdale, NJ: Lawrence Erlbaum, pp. 328–333.

Cummins, R. (1991) The role of representation in connectionist explanation of cognitive capacities. In W. Ramsey, S. Stich & D. Rumelhart (Eds), *Philosophy and Connectionist Theory.* Hillsdale, NJ: Lawrence Erlbaum.

Cussins, A. (1990) The connectionist construction of concepts. In M. Boden (Ed.) *The Philosophy of Artificial Intelligence.* Oxford: Oxford University Press.

Cybenko, G. (1989) Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems,* **2,** 303–314.

D'Amato, M. R. & Van Sant, P. (1988) The person concept in monkeys. *Journal of Experimental Psychology: Animal Behavior Processes,* **14,** 43–55.

Derthick, M. (1987) A connectionist architecture for representing and reasoning about structured knowledge. Research report CMU-BOLTZ-29. Pittsburgh, PA: Department of Computer Science, Carnegie-Mellon University.

Dowty, D. R., Wall, R. E. & Peters, S. (1981) *Introduction to Montague Semantics.* Dordrecht: Reidel.

Dretske, F. I. (1981) *Knowledge and the Flow of Information.* Cambridge, MA: MIT Press.

Dretske, F. I. (1986) Misrepresentation. In R. J. Bogdan (Ed.), *Belief: Form, Content and Function.* Oxford: Oxford University Press.

Dretske, F. I. (1988) *Explaining Behavior: Reasons in a World of Causes.* Cambridge, MA: MIT Press.

Duda, R. O. & Hart, P. E. (1973) *Pattern Classification and Scene Analysis.* Chichester: Wiley.

Dyer, M. G. (1990a) Finding lost minds. *Journal of Experimental and Theoretical Artificial Intelligence,* **2,** 329–339.

Dyer, M. G. (1990b) Intentionality and computationalism: Minds, Machines, Searle and Harnad. *Journal of Experimental and Artificial Intelligence,* **2,** 303–319.

Elman, J. L. (1988) Finding structure in time. Technical report CRL-8801. San Diego, CA: Centre for Research in Language, University of California.

Elman, J. L. (1990) Finding structure in time. *Cognitive Science,* **14,** 179–211.

Elman, J. L. (1991a) Distributed representation, simple recurrent networks, and grammatical structure. *Machine Learning,* **7,** 195–225.

Elman, J. L. (1991b) Incremental learning, or the importance of starting small. In *Proceedings from the 13th Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Lawrence Erlbaum, pp. 443–445.

Fahlman, S. E. (1988) An empirical study of learning speed in backpropagation networks. Technical report CMU-CS-88-192, Computer Science Department, Carnegie Mellon University, Pittsburgh, Pennsylvania.

Field, H. (1978) Mental representation. *Erkenntnis,* **13,** 9–61.

Finch, S. & Chater, N. (1992) Bootstrapping syntactic categories by unsupervised learning. In *Proceedings of the 14th Annual Meeting of the Cognitive Science Society,* Indiana University, Bloomington, July/August, pp. 820–825.

Fodor, J. A. (1975) *The Language of Thought.* New York: Thomas Crowell.

Fodor, J. A. (1981) The current status of the innateness controversy. In *Representations.* Cambridge, MA: MIT Press.

Fodor, J. A. (1984a) *Psychosemantics.* Unpublished manuscript.
Fodor, J. A. (1984b) Semantics, Wisconsin style. *Synthese,* **59,** 232–250.
Fodor, J. A. (1987) *Psychosemantics.* Cambridge, MA: MIT Press.
Fodor, J. A. (1990) *A Theory of Content and Other Essays.* Cambridge, MA: MIT Press.
Fodor, J. A. & McLaughlin, B. P. (1990) Connectionism and the problem of systematicity. Why Smolensky's solution doesn't work. *Cognition,* **35,** 183–204.
Fodor, J. A. & Pylyshyn, Z. W. (1988) Connectionism and cognitive architecture: a critical analysis. *Cognition,* **28,** 3–71.
Gold, E. (1967) Language identification in the limit. *Information and Control,* **16,** 447–474.
Goschke, T. & Koppelberg, D. (1991) The concept of representation and the representation of concepts in connectionist models. In W. Ramsey, S. Stich & D. Rumelhart (Eds) *Philosophy and Connectionist Theory.* Hillsdale, NJ: Lawrence Erlbaum.
Graf, H. P., Jackel, L. & Hubbard, W. E. (1988) VLSI implementation of a neural network. *Computer,* **21** (3), 41–51.
Grice, H. P. (1957) Meaning. *Philosophical Review,* **66,** 377–388.
Hampton, J. A. (1988) Overextension of conjunctive concepts: evidence for a unitary model of concept typicality and class inclusion. *Journal of Experimental Psychology: Learning, Memory and Cognition,* **14,** 12–32.
Hanson, S. J. & Burr, D. J. (1990) What connectionist models learn: learning and representation in connectionist networks. *Behavioral and Brain Science,* **13,** 471–518.
Hanson, S. J. & Kegl, J. (1987) PARSNIP: a connectionist network that learns natural language grammar from exposure to natural language sentences. In *Proceedings of the Eighth Annual Meeting of the Cognitive Science Society.* Hillsdale, NJ: Lawrence Erlbaum, pp. 106–117.
Harnad, S. (1989) Minds, machines and Searle. *Journal of Experimental and Theoretical Artificial Intelligence,* **1,** 5–25.
Harnad, S. (1990a) Lost in the hermeneutic hall of mirrors. *Journal of Experimental and Theoretical Artificial Intelligence,* **2,** 321–327.
Harnad, S. (1990b) The symbol grounding problem. *Physica D,* **42,** 335–346.
Harnad, S. (1992) Connecting object to symbol in modeling cognition. In A. Clark & R. Lutz (Eds), *Connectionism in Context.* Berlin: Springer-Verlag.
Harnad, S., Hanson, S. J. & Lubin, J. (1991) Categorical perception and the evolution of supervised learning in neural nets. Presented at the *1991 AAAI Symposium on Symbol Grounding: Problems and Practice.* Stanford University, March.
Hatfield, G. (1991) Representation in perception and cognition: connectionist affordances. In W. Ramsey, S. Stich & D. Rumelhart (Eds) *Philosophy and Connectionist Theory.* Hillsdale, NJ: Lawrence Erlbaum.
Haugeland, J. (1991) Representational genera. In W. Ramsey, S. Stich & D. Rummelhart (Eds) *Philosophy and Connectionist Theory.* Hillsdale, NJ: Lawrence Erlbaum.
Herrnstein, R. J., Loveland, D. H. & Cable, C. (1976) Natural concepts in pigeons. *Journal of Experimental Psychology: Animal Behaviour Processes,* **2,** 285–311.
Hinton, G. E. (1981) Implementing semantic networks in parallel hardware. In G. E. Hinton & J. A. Anderson (Eds), *Parallel Models of Associative Memory.* Hillsdale, NJ: Lawrence Erlbaum.
Hinton, G. E. (1986) Learning distributed representation of concepts. In *Proceedings of the Eighth Annual Meeting of the Cognitive Science Society.* Hillsdale, NJ: Lawrence Erlbaum, pp. 1–12.
Hofstadter, D. R. (1985) Waking up from the Boolean dream; or, subcognition as computation. In *Metamagical Themas: Questing for the Essence of Mind and Pattern.* New York: Viking.
Hopfield, J. J. (1982) Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences USA,* **79,** 2554–2558.
Horgan, T. & Tienson, J. (Eds) (1987) Connectionism and the philosophy of Mind. *The Southern Journal of Philosophy,* **26** (Suppl.).
Ingle, D. J. (1983) Brain mechanisms of visual localization by frogs and toads. In J. P. Ewert, A. Capranica & D. J. Ingle (Eds), *Advances in Vertebrate Neuroethology.* New York: Plenum Press.
Jelinek, F., Lafferty, J. D. & Mercer, R. L. (1990) Basic methods of probabilistic context free grammars. Technical report RC 16374 (72684). Yorktown Heights, NY: IBM.
Jordan, M. (1986) Serial order: a parallel distributed approach. Technical report 86040 San Diego, CA: Institute for Cognitive Science, University of California.
Kripke, S. (1972) Naming and Necessity. In D. Davidson & G. Harman (Eds), *Semantics of Natural Language.* Dordrecht: Reidel.
Kruschke, J. K. (1990) ALCOVE: a connectionist model of category learning. Research report 19. Bloomington, IN: Cognitive Science, Indiana University.

Kuhn, T. S. (1970) *The Structure of Scientific Revolutions,* 2nd edn. Chicago, IL: University of Chicago Press.

Lakoff, G. (1987) Cognitive models and prototype theory. In U. Neisser (Ed.), *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization.* Cambridge: Cambridge University Press.

Lea, S. E. G. (1984) In what sense do pigeons learn concepts? In H. Roitblat, T. G. Bever & H. S. Terrace (Eds), *Animal Cognition.* Hillsdale, NJ: Lawrence Erlbaum, pp. 263–276.

Lenat, D. B. (1982) AM: discovery in mathematics as heuristic search. In D. B. Lenat & R. Davies (Eds), *Knowledge-Based Systems in Artificial Intelligence.* New York: McGraw-Hill.

Lettvin, J. Y., Maturana, H., McCulloch, W. S. & Pitts, W. H. (1959) What the frog's eye tells the frog's brain. *Proceedings of the Institute of Radio Engineers,* **47,** 940–951.

Linsker, R. (1988) Self-organisation in a perceptual network. *Computer,* **21** (3), 105–117.

Mackay, D. J. C. (1991) Bayesian methods for adaptive models. PhD Thesis, California Institute of Technology.

Marr, D. (1982) *Vision.* San Francisco: Freeman.

McCauley, R. N. (1987) The role of theories in a theory of concepts. In U. Neisser (Ed.), *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization.* Cambridge: Cambridge University Press.

McClelland, J. L. & Kawamoto, A. H. (1986) Mechanisms of sentence processing. In J. L. McClelland & D. E. Rumelhart (Eds), *Parallel Distributed Processing,* Vol. 2. Cambridge, MA: MIT Press.

McClelland, J. L. & Rumelhart, D. E. (Eds) (1986) *Parallel Distributed Processing: Explorations in the Microstructures of Cognition,* Vol. 2. *Psychological and Biological Models.* Cambridge, MA: MIT Press.

McGinn, C. (1989) *Mental Content.* Oxford: Basil Blackwell.

Medin, D. L. & Shoben, E. J. (1988) Context and structure in conceptual combination. *Cognitive Psychology,* **20,** 158–190.

Medin, D. L. & Wattenmaker, W. D. (1987) Category cohesiveness, theories, and cognitive archeology. In U. Neisser (Ed.), *Concepts and Conceptual Development: Ecological and Intellectual Factors in Categorization.* Cambridge: Cambridge University Press.

Millikan, R. G. (1984) *Language, Thought and Other Biological Categories.* Cambridge: Cambridge University Press.

Millikan, R. G. (1986) Thought without laws: cognitive science with content. *Philosophical Review,* **95,** 47–80.

Minsky, M. L. & Papert, S. (1969) *Perceptrons: An Introduction to Computational Geometry.* Cambridge, MA: MIT Press.

Moody, J. E. (1992) The effective number of parameters: an analysis of generalization and regularization in nonlinear learning systems. In J. E. Moody, S. J. Hanson & R. P. Lippmann (Eds), *Advances in Neural Information Processing Systems,* Vol. 4. San Mateo, CA: Morgan Kaufmann.

Murphy, G. L. & Medin, D. L. (1985) The role of theories in conceptual coherence. *Psychological Review,* **92,** 289–316.

Newell, A. & Simon, H. A. (1976) Computer science as empirical inquiry. *Communications of the ACM,* **19,** 113–126. Reprinted in M. Boden (Ed), *The Philosophy of Artificial Intelligence.* Oxford: Oxford University Press.

Niklasson, L. & Sharkey, N. E. (1992) Connectionism and the issues of compositionality and systematicity. Presented at the *1992 EMCSR Symposium on Connectionism and Cognitive Processing,* University of Vienna, April, 1992.

Norris, D. G. (1990) A dynamic-net model of human speech recognition. In G. Altmann (Ed.), *Cognitive Models of Speech Processing: Psycholinguistic and Cognitive Perspectives.* Cambridge, MA: MIT Press.

Oaksford, M., Chater, N & Stenning, K. (1990) Connectionism, classical cognitive science and experimental psychology. *AI and Society,* **4,** 73–90.

Oaksford, M. & Chater, N. (1991) Against logicist cognitive science. *Mind and Language,* **6,** 1–38.

Osherson, D., Stob, M. & Weinstein, S. (1986) *Systems That Learn: An Introduction to Learning Theory for Cognitive and Computer Scientists.* Cambridge, MA: MIT Press.

Piatelli-Palmarini (Ed.) (1980) *Language and Learning: The Debate between Jean Piaget and Noam Chomsky.* Cambridge, MA: Harvard University Press.

Pinker, S. (1979) Formal models of language learning. *Cognition,* **7,** 217–283.

Pinker, S. (1984) *Language Learnability and Language Development.* Cambridge. MA: Harvard University Press.

Pinker, S. & Prince, A. (1988) On language and connectionism. *Cognition,* **28,** 73–195.

Pollack, J. B. (1988) Recursive auto-associative memory: devising compositional distributed representa-

tions. In *Proceedings of the 10th Annual Meeting of the Cognitive Science Society.* Hillsdale, NJ: Lawrence Erlbaum, pp. 33–39.

Pollack, J. B. (1990) Recursive distributed representations. *Artificial Intelligence,* **46,** 77–105.

Pylyshyn, Z. (1984) *Computation and Cognition: Toward a Foundation for Cognitive Science.* Cambridge: Cambridge University Press.

Quine, W. (1960) *Word and Object.* Cambridge, MA: MIT Press.

Ramsey, W., Stich, S. & Rummelhart, D. (Eds) (1991) *Philosophy and Connectionist Theory.* Hillsdale, NJ: Lawrence Erlbaum.

Rumelhart, D. E. & McClelland, J. L. (Eds) (1986) *Parallel Distributed Processing: Explorations in the Microstructures of Cognition,* Vol. 1. *Foundations.* Cambridge, MA: MIT Press.

Rumelhart, D. E., Smolensky, P., McClelland, J. L. & Hinton, G. E. (1986) Schemata and sequential thought processes in PDP models. In J. L. McClelland & D. E. Rumelhart (Eds), *Parallel Distributed Processing,* Vol. 2. Cambridge, MA: MIT Press.

Salmon, N. U. (1982) *Reference and Essence.* Oxford: Basil Blackwell.

Schiffer, S. (1987) *Remnants of Meaning.* Cambridge, MA: MIT Press.

Schneider, W. (1987) Connectionism: is it a paradigm shift for psychology? *Behaviour, Research Methods, Instruments, & Computers,* **19,** 73–83.

Schurg-Pfeiffer, E. & Ewert, J. P. (1981) Investigation of neurons involved in the analysis of Gestalt prey features in the frog *Rana temporaria. Journal of Comparative Physiology,* **141,** 139–158.

Searle, J. R. (1980) Minds, brains and programs. *Behavioral and Brain Sciences,* **3,** 417–424. Reprinted in M. Boden (Ed.), *The Philosophy of Artificial Intelligence.* Oxford: Oxford University Press.

Searle, J. (1990) Is the brain's mind a computer program? *Scientific American,* **262,** 20–25.

Servan-Schreiber, D., Cleeremans, A. & McClelland, J. L. (1989) Learning sequential structure in simple recurrent networks in D. Touretzky (Ed.), *Advances in Neural Information Processing Systems,* Vol. 1, Palo Alto, CA: Morgan Kaufman, pp. 643–653.

Servan-Schreiber, D., Cleeremans, A. & McClelland, J. L. (1991) Graded state machines: the representation of temporal contingencies in simple recurrent networks. *Machine Learning,* **7,** 161–193.

Sharkey, N. E. (1991) Connectionist representation techniques. *AI Review,* **5,** 143–167.

Sharkey, N. E. (1992) The causal role of the constituents of superpositional representations. Presented at the *1992 EMCSR Symposium on Connectionism and Cognitive Processing,* University of Vienna, April 1992.

Shillcock, R., Levy, J. & Chater, N. (1991) A connectionist model of word recognition in continuous speech. In *Proceedings from the 13th Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Lawrence Erlbaum, pp. 340–345.

Smolensky, P. (1987) The constituent structure of connectionist mental states: a reply to Fodor and Pylyshyn. *Southern Journal of Philosophy,* **26,** 137–159.

Smolensky, P. (1988) On the proper treatment of connectionism. *Behavioral and Brain Sciences,* **11,** 1–74.

Sopena, J. M. (1991) ERSP: a distributed connectionist parser that uses embedded sequences to represent structure. Technical report UB-PB-1-91. Barcelona: Departament de Psicologia Bàsica, Universitat de Barcelona.

Stampe, D. (1977) Toward a causal theory of linguistic representation. In P. French, T. Euhling & H. Wettstein (Eds), *Midwest Studies in Philosophy 2,* Minneapolis: University of Minnesota Press.

Stich, S. (1983) *From Folk Psychology to Cognitive Science.* Cambridge, MA: MIT Press.

Stolcke, A. (1991) Syntactic category formation with vector space grammars. In *Proceedings from the 13th Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Lawrence Erlbaum, pp. 908–912.

Touretzky, D. S. & Hinton, G. E. (1985) Symbols among the neurons: details of a connectionist inference architecture. In *Proceedings of the Ninth International Joint Conference on Artificial Intelligence,* University of California at Los Angeles, Los Angeles, pp. 238–243.

van Gelder, T. (1990) Compositionality: a connectionist variation on a classical theme. *Cognitive Science,* **14,** 355–384.

van Gelder, T. (1991) What is the 'D' in 'PDP'? A survey of the concept of distribution. In W. Ramsey, S. Stich & D. Rumelhart (Eds), *Philosophy and Connectionist Theory.* Hillsdale, NJ: Lawrence Erlbaum.

van Gelder, T. (1992) Classical questions, radical answers: connectionism and the structure of mental representations. In T. Horgan & J. Tienson (Eds), *Connectionism and the Philosophy of Mind.* Dordrecht: Kluwer.

Winston, P. H. (1975) Learning structural descriptions from examples. In P. H. Winston (Ed.), *The Psychology of Computer Vision.* New York: McGraw-Hill.