Cognition, 48 (1993) 223–242

Neural networks, nativism, and the plausibility of constructivism

Steven R. Quartz

Departments of Cognitive Science and Philosophy, University of California, La Jolla, CA 92186-5800, USA and Computational Neurobiology Laboratory, The Salk Institute for Biological Studies, P.O. Box 85800, San Diego, CA 92186-5800, USA

Received March 11, 1992, final version accepted May 12, 1993

Abstract

Recent interest in PDP (parallel distributed processing) models is due in part to the widely held belief that they challenge many of the assumptions of classical cognitive science. In the domain of language acquisition, for example, there has been much interest in the claim that PDP models might undermine nativism. Related arguments based on PDP learning have also been given against Fodor's anti-constructivist position – a position that has contributed to the widespread dismissal of constructivism. A limitation of many of the claims regarding PDP learning, however, is that the principles underlying this learning have not been rigorously characterized. In this paper, I examine PDP models from within the framework of Valiant's PAC (probably approximately correct) model of learning, now the dominant model in machine learning, and which applies naturally to neural network learning. From this perspective, I evaluate the implications of PDP models for nativism and Fodor's influential anti-constructivist position. In particular, I demonstrate that, contrary to a number of claims, PDP models are nativist in a robust sense. I also demonstrate that PDP models actually serve as a good illustration of Fodor's anti-constructivist position. While these results may at first suggest that neural network models in general are incapable of the sort of concept acquisition that is required to refute Fodor's anti-constructivist position, I suggest

The author was supported by a Postgraduate Fellowship from the Natural Sciences and Engineering Research Council of Canada and a McDonnell-Pew Center for Cognitive Neuroscience Fellowship. The author would like to thank P. Dayan, P.R. Montague, and P.S. Churchland for helpful discussion. The comments of two anonymous referees were also helpful.

Correspondence to: S.R. Quartz, Departments of Cognitive Science and Philosophy, University of California, La Jolla, CA 92186-5800, USA.

that there is an alternative form of neural network learning that demonstrates the plausibility of constructivism. This alternative form of learning is a natural interpretation of the constructivist position in terms of neural network learning, as it employs learning algorithms that incorporate the addition of structure in addition to weight modification schemes. By demonstrating that there is a natural and plausible interpretation of constructivism in terms of neural network learning, the position that nativism is the only plausible model of acquisition can no longer be defended. Indeed, I briefly discuss a number of learning-theoretic reasons indicating that constructivist models so characterized uniquely possess a number of important learning characteristics.

1. Introduction

One of the few detailed theories of cognitive development is the constructivist model. Unlike the heavy burden nativism places on genetic mechanisms to account for specific brain structures, constructivism regards the emergence of these structures as the outcome of the interaction between developmental mechanisms and the structure of the environment in which the organism is embedded – an interaction that developmental neurobiology is coming to appreciate (e.g., Hockfield & Kalb, 1993; Shatz, 1990).[1] Despite the appeal of many of its features, there has been a widespread dismissal of constructivism, due in part to Jerry Fodor's (1975, 1980, 1981) arguments against its coherence. For example, Pinker (1984) cites Fodor's arguments as grounds for dismissing constructivist models of development in the case of language acquisition (for an extended discussion, see Piattelli-Palmarini, 1980). Indeed, the nativist assumptions that dominate cognitive science are in some measure the result of the widely held belief that there are no plausible alternative models of acquisition.

With the rise of PDP (parallel distributed processing) networks in recent years, an intriguing possibility is that these models may offer a view of learning that refutes the assumptions underlying positions such as nativism and the anti-constructivist one mentioned above.[2] Although this claim has been most intensely scrutinized in terms of the language acquisition properties of PDP models and their implications for nativism (e.g., Pinker & Prince, 1988; Rumelhart & McClelland, 1986; for a review see Bates, 1992), related arguments have also

[1] The recent adoption of so-called selectionist models (e.g., Piattelli-Palmarini, 1989; Lightfoot, 1991, 1992) may be seen as an attempt to maintain the nativist position without the commitment to explicit genetic encodings. In Quartz (1993a) I suggest that the learning-theoretic properties of selectionist models render them infeasible.

[2] I use PDP models and connectionism interchangeably to refer to fixed feedforward neural networks where learning is construed as an adaptive changing of node functions, such as gradient descent minimization of a scalar error function.

been given against the anti-constructivist position (e.g., Chater & Oaksford, 1990). Both the anti-nativist arguments and those directed against Fodor's anti-constructivist position share the belief that PDP models learn by creating internal representations rather than having these representations built in – a popular conception of connectionist learning. A fundamental limitation of these arguments, however, is that the characterization of PDP learning they employ is not a rigorous one. Therefore, the force of the claim that PDP models both circumvent nativism and establish the plausibility of constructivism is not clear in the absence of a more rigorous treatment of the properties of this class of models.

By introducing some elementary learning-theoretic notions, in what follows I will demonstrate that, once characterized from a more formal perspective, standard PDP models are in fact nativist, and, rather than undermine Fodor's arguments, they actually serve as a good illustration of his anti-constructivist position. Indeed, contrary to a number of claims, PDP models are typically more highly constrained than classical architectures. Does this result imply that the nativist, anti-constructivist arguments typified by Fodor's position are corroborated, as they now appear to hold for the distinct style of representation that PDP models offer? By further considering learning from this formal perspective, the answer to this question will be negative, as I will suggest that Fodor's position against constructivism is undermined by considering an alternative form of learning in neural networks. This alternative form of learning is a natural interpretation of the constructivist position in terms of neural network learning, as it employs learning algorithms that incorporate the addition of structure.

As I describe in more detail below, Fodor's argument against constructivism depends on first defining some quantitative measure of a conceptual structure's complexity and then showing that this measure cannot increase over time. But, by employing a quantitative measure of a conceptual structure's complexity from this alternative model of learning in neural networks, I will demonstrate that there is a natural sense in which Fodor's argument against constructivism dissolves as an artifact of his choice of this measure. Although my main aim will be in demonstrating the plausibility of the constructivist position, since Fodor's arguments are directed at this question, I will also briefly suggest that there are a number of learning-theoretic reasons for supposing that constructivist models so characterized uniquely possess a number of important learning characteristics that fundamentally differ from standard PDP models.

2. Fodor's arguments

Before considering the possible relevance of connectionist models to the question of constructivism's plausibility, it is worthwhile briefly to spell out Fodor's arguments. With the premises of Fodor's arguments made explicit, it will then be possible to consider the conditions that must be satisfied by a candidate learning model to be a constructivist system and, specifically, whether connectionist models provide such an account.

Fodor's argument against constructivism depends on showing that the notion of concept learning in general is confused and that, on closer inspection, there could not be any such phenomenon. Roughly, in order for there to be a sense in which a cognitive system could acquire new concepts, there must be a well-defined sense in which the system's conceptual resources increase over time. As this is exactly the Piagetian notion of stages of conceptual development (Piaget, 1954), in which successive stages are richer in their representational power than previous ones, Fodor's argument is aimed at constructivism as a special case of this more general argument against concept acquisition.

From Fodor (1980), we may formulate the following conditions on concept acquisition:

(1) The concept must originally lie outside the domain of the learner's conceptual repertoire (initial state).
(2) Through some process, the learner must come to acquire the concept.
(3) The learner must learn the truth conditions for that concept.

For (3), there is a well-developed sense of learning as inductive inference; however, regarding (2), Fodor (1981) contends that there is no available theory of the source of our inductive hypotheses, but that theories of concept learning simply presuppose their availability. Fodor's arguments have a further aim: not only is there no theory of concept acquisition as a contingent fact about the state of cognitive psychology, but there could not be such a theory. We are thus led to nativism by a dismissal of any cogent alternative.

In what sense could it be said that there could never be a theory of concept acquisition? To make this question more precise, it is necessary to adopt a quantitative measure of the representational power of a cognitive structure, as this measure will index conceptual expressiveness and must therefore increase over time with the complexity of the system. Fodor (1980) considers such a measure in terms of the logic instantiated by a cognitive system, which we may formulate as the following condition:

> a system at some time$_t$ is a more powerful structure than at time$_{t-i}$ only if the set of truths that could be expressed by the logic the system instantiates at time$_t$ is larger than the set of truths that could be expressed by the logic the system instantiates at time$_{t-i}$ (i.e., the set of truths of the logic$_{t-i}$ is a proper subset of the truths of the logic at time$_t$).

Now that this condition for acquiring novel concepts is explicit, it is possible to consider whether it can be satisfied in general. To show that accepting this picture entails that there can be no acquisition of a more powerful logic in terms of learning as inductive inference, Fodor (1980) considers the case of moving from propositional logic at stage$_i$ to first-order quantificational logic at stage$_{i+1}$. To learn quantificational logic, the system must learn the truth conditions of its formulae, such as "'(X)Fx' is true if and only if . . . ," as condition (3) specifies. Yet, this requires that the system have at stage$_i$ the conceptual apparatus sufficient to represent this truth-conditional statement. But, by definition, a system instantiating propositional logic cannot represent this statement. The conclusion, according to Fodor (1980, p. 149), is:

> . . . there literally isn't such a thing as the notion of learning a conceptual system richer than the one that one already has; we simply have no idea of what it would be like to get from a conceptually impoverished to a conceptually richer system by anything like a process of learning. Thus there is an important sense in which the nativist hypothesis is the only one in the field . . .

In evaluating this argument, it is important to point out that Fodor's argument depends on establishing some fairly rigorous sense of representational power in order to establish a quantitative notion of the change that is supposed to occur between stages of development. For this, Fodor employs the notion of a logic, as we saw above, which establishes this quantitative measure as the set of truths a logic expresses. However, the choice of logic for this measure has no privileged status in itself. The notion of a logic does lend itself naturally to a quantitative treatment of representational power or expressiveness, especially under a view that regards cognition to depend essentially on a language of thought (LOT), as the representational units in such a model are formal strings or formulae with a recursive syntax (Pylyshyn, 1984). Yet its use is not mandated as the only possible quantitative measure, even within LOT models. It is this point that makes connectionist models interesting to the question of the plausibility of constructivist models, as PDP models have been offered as an alternative to LOT models of cognition. In particular, what makes connectionism of potential relevance to this question is the possibility that there exists an appropriate quantitative measure of representational power from within this domain of research that will help remove these obstacles to concept acquisition.

3. PDP models are inadequate to refute Fodor's anti-constructivist position

As I mentioned above, Chater and Oaksford (1990) suggest that the learning characteristics of PDP models refute Fodor's argument against constructivism; they state (p. 101):

> By contrast [to PDP models], standard learning models cannot develop new structures . . . since Classical learning is just hypothesis generation and confirmation. Everything that can be learnt must be represented innately . . . PDP promises a theory of learning which sidesteps these difficulties.

Is it the case that PDP models sidestep the difficulties associated with classical learning models? In light of the above discussion, this now reduces to the question of whether PDP models offer an account of concept acquisition and whether an alternative measure of representational power can be derived from this model that makes sense of increases to a system's conceptual resources over time.

To consider the questions of concept acquisition and nativism from within the domain of connectionist research, we now require a general conception of connectionist network learning. In particular, as we are interested in evaluating the capacities of connectionist network learning and in defining such notions as the initial state of a network, we require a fairly rigorous model. The Valiant (1984) framework has recently been applied to neural network learning (e.g., Abu-Mostafa, 1989; Baum, 1989; Baum & Haussler, 1989) and provides this general conception, as I outline below.

To see the utility of the application of the Valiant model to neural network learning, it is useful first to consider the problem of feasibility in learning. In particular, it was seen early on that a general learner (e.g., Gold, 1967), while perhaps capable of learning in the limit, was not capable of learning in feasible time. Instead, as Pinker (1979) notes in the case of language acquisition, such a general learner may have a test on the order of 10^{100} possible grammars even in an extremely simplified case - a computation that could never actually be performed. Gold's learner was so slow because it adopted a general strategy whereby it simply enumerated an entire class of grammars and hypothesized each element in turn until it reached the target grammar. Within the Chomsky hierarchy of languages, primitive recursive languages are the highest class that is learnable by such a strategy since they are the highest decidable class. Although this guarantees convergence its practical implications are severely limited because of the vast search that may be required. A fundamental limitation of Gold's paradigm, then, was that it did not address the problem of learning in feasible time by incorporating complexity constraints into its model of learning - considerations that are of prime importance to learning both in natural systems and in large real-world applications.

The most notable example of applying complexity considerations to learning is Valiant's (1984) PAC (probably approximately correct) model of learning, which has now become the dominant model in machine learning (e.g., Natarajan, 1991). Valiant's original study concerned the learning in polynomial time (polynomial in the number of arguments of the function) of a Boolean function f in a class of Boolean functions F on the domain Σ^* from examples chosen according to some arbitrary but fixed probability distribution D, where Σ is the Boolean alphabet $\{0, 1\}$ and Σ^* is the set of strings of finite length on Σ. On this model, the learner has access to a subroutine EXAMPLE that provides positive and negative examples, where each example is a feature vector for which $f(\vec{x}) = 1$ or 0 that are drawn according to the probability distribution D. A condition of Valiant's model is that the probability distribution D that is employed in the training case likewise be employed in the test cases. A function is then said to be learned by some algorithm just in case when supplied with examples of f drawn from D, the algorithm constructs with probability $1-\delta$ a hypothesis g belonging to a representation class G such that $g(\vec{x}) = f(\vec{x})$ with probability at least $1-\varepsilon$ (δ and ε are defined below).

An important insight into learning from this model was that feasible learning – learning that is achieved within some realistic time bounds – was seen to require (at least) a significant restriction of the possible conjectures that the learner must evaluate (see Blumer, Ehrenfeucht, Haussler & Warmuth, 1987). A step towards achieving this in the Valiant model is to significantly restrict the class of concepts that the system may represent by delimiting in advance the hypothesis space or representation class that the learner may employ. The question of feasibility may then be addressed by determining the complexity of learning various functions from this class. In particular, a hypothesis space will be polynomially learnable just in case the number of examples required is polynomial as a function of n, $1/\varepsilon$, $1/\delta$, where n is a length parameter on the examples, and a consistent hypothesis can be found in G in time polynomial in n, $1/\varepsilon$, $1/\delta$. In addressing these issues, Valiant's model thus shifts the main emphasis of the learning problem from what is in principle learnable to what is learnable from some restricted representation class in feasible time. This relativization of learning to some specified representation class has a direct application to PDP models, as I consider next.

Valiant's model may be applied to PDP networks by considering an arbitrary, feedforward architecture $\mathcal{G}$ (see Baum & Haussler, 1989), a class of networks that share the same directed acyclic graph G (roughly, the pattern of connections between processing nodes), which may serve generally to define the learning-theoretic characteristics of the class of connectionist models. Associated with each node (excluding input nodes) is a node function set, F_i. Each individual member $\mathfrak{g}$ belonging to $\mathcal{G}$ may be constructed by choosing a particular f_i from F_i for each appropriate node (see Baum & Haussler, 1989). $\mathcal{G}$ may therefore be regarded as the set of all those networks that may be realized by setting the function that each node may compute to some value while holding constant the pattern of connectivity between nodes.

We can now see how Valiant's model provides a general characterization of connectionist network learning. Specifically, a fixed feedforward architecture represents a certain class of concepts that will be determined by the connectivity between nodes and the functions that each node may perform. A feedforward architecture $\mathcal{G}$ will thus be identified with the representation class G in Valiant's model. Roughly, as learning proceeds, this class of concepts is reduced by some error-correction method until the network settles on some element g of G that approximates the target function. As in Valiant's model, the end state of the network need not be identical to the target function; instead, it need only approximate it within certain parameters. What is important, however, is that the probability distribution that was employed in training the network likewise be employed in evaluating the network's ability to generalize after training. Valiant's framework thus contains the same definition of generalization that is used in neural network research and it allows for the network to make a fraction ε of incorrect predictions and to fail with some probability δ, features which are necessary since the examples are chosen probabilistically. δ is a confidence parameter and ε is an error parameter, where in successful learning the algorithm is $(1-\delta)$ confident that its error is at most ε. Among the most important applications of this framework to neural networks has been the obtaining of bounds on the relation among network size, training sample size, and generalization ability (Baum & Haussler, 1989).

Given the applicability of Valiant's model to learning in feedforward neural networks, we are now in a position to consider from this perspective both whether connectionist networks solve Fodor's problem of concept acquisition, as Chater and Oaksford (1990) suggest, and whether these models escape nativism.

Regarding Fodor's arguments against concept acquisition, we may now consider whether PDP models satisfy condition (2) above, as it is this condition that appeared to be problematic for classical models of learning. As condition (2) specifies as a matter of definition, in order for a network to be capable of concept acquisition, it would be required to come to represent some concept h that is not an element of its initial state. In order to evaluate this condition, we thus require a characterization of the initial state of a network. Here the application of the Valiant framework to neural networks proves useful. In the Valiant model, the initial state is represented as a hypothesis space G, from which the end state, which approximates the target function within some tolerance parameters, is chosen by the restrictive pressures of the input space. The application of this model to neural networks resulted in the identification of the feedforward architecture $\mathcal{G}$ with the hypothesis space G, which may therefore be considered the initial state of a network.

As we have now identified the initial state of an architecture $\mathcal{G}$ with the hypothesis space G, the question of the plausibility of concept acquisition reduces to whether such an architecture may come to represent some concept not belonging to G. It is now clear, though, that connectionist models in general are incapable of representing any concept h that is not an element of its initial state G. This follows from consideration of the demands of feasible learning, which PDP models implicitly incorporate, and from the structural features of this class of models. As the complexity considerations of the Valiant model illustrate, feasible learning requires that learning be relative to some restricted hypothesis space that the learner employs. This we identified with the initial state of the network, G, which may be roughly regarded as the innate knowledge of the system or as the inductive bias, as it is known in the machine-learning literature.

However, the combination of both feasibility demands that restrict the initial state to the representation class G and the structural features of connectionist models that identify this class with a fixed feedforward architecture entails that no concept lying outside the initial state may be represented by the model. The reason for this is that the only free parameters of a PDP model are the weights (and thresholds), which may vary according to some error-correction method. But, the effect of learning is to reduce the possible concepts that the system may represent, as learning reduces the set of elements of G that are consistent with the training examples to a proper subset of G (and ultimately to a particular element of G). Therefore, condition (2) cannot be satisfied by connectionist models, and so these models do not offer the appropriate account of concept acquisition.

A response to this argument may be to point out that PDP models are nonetheless capable of coming to represent some concept that is not explicitly represented in the initial state. Yet, Fodor is not committed to the extreme view that every concept must be represented explicitly in the initial state. According to Fodor (1981), only a subset of possible primitive concepts comes to be represented explicitly, namely those that are "triggered" by appropriate environmental features. While avoiding identifying "triggering" (which appears never to have been characterized rigorously as an acquisition mechanism) with neural network learning, it is clear what Fodor's general position corresponds to in terms of connectionist network learning. The particular hypothesis g that is arrived at need not be represented explicitly in the initial state. Indeed, typically it would not be since that would make learning trivial. Instead, arriving at the end state g is derivable from some combination of weights from the initial state through the representation class G to g. Yet, any element that belongs outside G, the initial state, cannot be represented by the network, despite any possible configuration of the weights.

Put another way, Fodor's argument is a challenge to show how a cognitive structure may increase its complexity over time as a function of learning. In the case of connectionist networks, any increase in the complexity measure will be dependent on the size of G – the class of concepts a network may represent. Yet, by definition G cannot increase over time, since we identified G with the fixed feedforward architecture $\mathcal{G}$, and, hence, any complexity measure that is dependent on $\mathcal{G}$ will be time-invariant.

4. PDP models are nativist

The results of the application of Valiant's model to neural networks for these debates are in fact stronger than those of the last section. So far, I have outlined only how it is that PDP models are incapable of coming to represent some concept that lies outside their initial state. Rather than learn by creating internal representations, as they have often been characterized, they actually learn by eliminating elements of an *a priori* defined hypothesis space to those that are consistent with the training examples. The Valiant model, however, also illustrates the stronger result that PDP models are nativist in a robust sense. These are distinct claims, as it may be the case that while PDP models cannot acquire concepts lying outside their initial state, the representation class that is represented innately by a feedforward architecture may be so general that the network may be applicable across a number of domains. If this were the case, then PDP models would not need to build in domain-specific knowledge and could be regarded as implementing general learning strategies. In the context of applications to complex problem domains, however, this is an unlikely possibility. From within Valiant's model, general arguments follow from complexity considerations that show that the initial state of a PDP network must be an exponentially small subset of all possible concepts, U, in order for the network to feasibly learn (see Blumer et al., 1987). Hence, as Dietterich (1990) points out, the probability that some particular network may PAC learn on a hypothesis chosen at random from all possible hypotheses U is vanishingly small.

The abstract characterization afforded by the application of Valiant's model thus explains why the choice of an architecture is such an important factor in connectionist learning (see Hertz, Krogh & Palmer, 1991, Ch. 6), as it is the choice of this hypothesis space. It also illustrates why so much of neural network research is devoted to automating the search for appropriate network architectures for particular problem domains and to finding useful heuristics to guide this selection. Viewing the problem from a statistical perspective, Geman, Bienenstock, & Doursat (1992, p. 45) thus conclude that in real-world applications the only means of finding practical solutions to large neural modeling problems is to prewire the important generalizations and that finding the appropriate initial state of a network so that it lies close to the target function is the fundamental problem of neural network research. From a different framework, Dente and Mendes (1992) reach similar negative conclusions against general learning strategies for PDP models. Although it is sometimes remarked that PDP models are *tabula rasa* learners in virtue of having random initial weights, the effect of initial weight randomization is really just to place the network somewhere in an *a priori* defined hypothesis space, which is necessary for applications.

It is in part because PDP models represent such a highly constrained hypothesis space that they display favorable learning characteristics. Indeed, although connectionist networks are sometimes regarded as more powerful computationally than Turing-based models, the favorable learning characteristics of connectionist models actually stem from their structural restrictions that in limited domains result in fast learning. However, this comes at a cost: the network must be tailored to the task and will with a high likelihood fail to learn in some unanticipated problem domain. Therefore, the notion that a generalpurpose network architecture may be found is untenable. It should be noted, however, that this result against general learning is a property of the problem of learning by examples from a defined hypothesis space and not of a particular algorithm, and so any approach – whether Turing-based or PDP-based – that views learning as the search through a hypothesis space will be constrained by these results.

For these reasons, the above argument does not amount to merely defining some notion of representational capacity and then claiming that a network cannot exceed this capacity. By definition, a system cannot exceed its representational capacity, and so this would be a trivial position. Rather, as the Valiant model illustrates, the *a priori* representation class or hypothesis space that a learner employs represents a specific and exponentially small subset of all possible concepts, of which an acceptable approximation of the target function is a member, as a necessary condition of feasible learning. This framework maps directly onto PDP models and identifies this hypothesis space with the structural features of a network architecture. The characterization of the initial state of a fixed architecture as a hypothesis space on the Valiant model is therefore stronger than some general measure of representational capacity. Thus, failure to learn may not be simply the result of inadequate resources stemming from circuit complexity considerations, in which the complexity of the target function exceeds the representational capacity of the network, but in some cases may be due to the selection of the wrong architecture as the representation of the hypothesis space.

In summary, to return to Fodor's argument against concept learning and its relevance to PDP models, the application of the Valiant framework to PDP models illustrates that these models have built into their architecture a highly restricted hypothesis space that contains the target function, or at least an acceptable approximation to it. Further, the fact that the architecture that represents this hypothesis space is fixed entails that no concept lying outside this hypothesis space may ever be represented by the network. Training by examples on such an architecture is, therefore, simply learning the truth conditions for a concept that the network already has available to it.

5. Alternative network models demonstrate the plausibility of constructivism

In the previous section, standard PDP models were shown to be strongly nativist, contrary to the popular conception of PDP learning as learning by creating internal representations. Given their strong nativism and their structural properties, it quickly followed that they did not offer an account of concept acquisition, in Fodor's sense of the term. Although this result may at first suggest that neural networks are in general incapable of offering an account of concept acquisition, within the framework of the last section it is straightforward to identify the features of standard PDP models that are responsible for this restriction on learning and to consider the nature of learning in their absence.

Informally, what is required to refute Fodor's position is some sense in which a system may increase it representational power – defined as the set of concepts it may express – as a function of learning. From the application of Valiant's model to neural network learning, it became apparent that the class of concepts G a network represents is identified with a fixed feedforward architecture $\mathcal{G}$ and that these fixed structural features restrict the possible concepts that may be represented to elements of that hypothesis space. But, since the architecture of a network is identified with the class of concepts it may represent, this suggests that a network with the ability to alter its architecture in appropriate ways as a function of learning will be capable of extending its representation class beyond its initial state and will therefore be capable of acquiring novel concepts. As I will consider in more detail in section 6, the characterization of a system that may add structure through learning is not exclusive to neural networks since Turing-based algorithms may be formulated that perform a similar operation. Yet, since neural networks effectively collapse the distinction between structure and function, in contrast to functionalist-based symbolic models, neural networks naturally allow for the investigation of how structural modifications may have significant functional consequences at the level of the representations that are supported by such a system.

Two conditions must be satisfied by such a system: (1) the addition of structure must be non-trivial by being describable as a process of learning; (2) it must offer an appropriate quantitative measure for progressive increases in representational power. I consider these two conditions below in the context of constructive neural networks, systems that alter their architecture by adding computational units as well as by modifying connection strengths, a natural interpretation of constructivism in terms of neural network learning.[3] Of course, such a learning mechanism would be of limited value if it did not possess powerful learning-theoretic characteristics. Although my main aim is in establishing the plausibility of constructivist learning, since this is what Fodor's arguments are aimed at, I will briefly consider these learning-theoretic properties in section 7.

6. Incorporating structural modifications

Constructivist models are examples of systems that incorporate the principle of non-stationarity, as it is known in the theory of computation, namely the property

[3] Although constructivist networks add units, it is reasonable to assume that synapses are the basic computational units of the brain (see Shepherd, 1990) and that constructivist processes will be identified with activity-dependent synaptogenesis in neurobiological systems. For a more extensive discussion of constructivist learning in neurobiological terms, see Quartz (1993b).

that a system may make changes to its underlying mechanism in addition to changes to the data structures that mechanism supports (for a discussion of this principle in terms of learnability, see Pinker, 1981). An objection to the inclusion of this principle, and hence to constructivism, is that it trivializes an explanation of learning or development since arbitrary processes may transform the learning system into the target state (e.g., Pinker, 1984). This objection amounts to the challenge to show that the process by which elements are added to a system are not arbitrarily related to the content of the learning episode that invokes them, but can be described as a cognitive process of learning, as condition 1 stipulated above.

The best response to this objection lies in the existence of a number of constructivist algorithms where the addition of new units is a principled one and which confer convergence properties on the networks, although a detailed analysis of these is beyond the scope of this paper (see, for example, Fahlman, 1991; Fahlman & Lebiere, 1990; Frean, 1990; Nadel, 1989; Wynne-Jones, 1993).[4] It should also be noted that one of the best examples of plasticity in neural systems, Hebbian plasticity – a neural implementation of associative learning (reviewed in Sejnowski & Tesauro, 1989) – was originally proposed by Hebb (1949) as a growth algorithm underlying associative learning and the activity-dependent construction of cell assemblies. Hebb's proposals may be viewed as a learning algorithm for the activity-dependent construction of neural circuits that has a clear interpretation in terms of the associative conditions that are identified with processes of learning, although discussion of this point is not possible here (see Quartz, 1993b). In fact, the use-dependent addition of structure underlying learning is one of the best documented types of neural plasticity, both in development and in the mature state as well (e.g., Black, Isaacs, Anderson, Alcantara, & Greenough, 1990; reviewed in Greenough & Bailey, 1988).

Despite the charge of trivialization, inclusion of changes to the underlying architecture is not in itself any more liable to introduce arbitrary changes than are other forms of plasticity. For example, this objection may also be directed at weight modification schemes in that it is equally possible to implement arbitrary rules for weight change that are not related to the nature of the input in any principled manner. What makes non-stationarity potentially problematic is not that it introduces arbitrary changes to a system, but that it incorporates qualitative changes to the learning mechanism that may be analytically difficult to evaluate. However, as I consider below, these qualitative changes may be of central importance in theories of development and may dramatically alter the learning capacities of developing systems.

[4] For example, Wynne-Jones's (1993) constructive algorithm uses principal component analysis to determine the area of the function space that is least well covered by the hidden units and then splits those units along the direction of maximal variance to better cover the space.

Non-stationarity has been excluded in a number of developmental studies in favor of the methodological principle known as the continuity hypothesis – that learning in development is fundamentally like learning in the mature state in terms of both the underlying processes and structures (Pinker, 1984; McNamara, 1982). Although the continuity hypothesis is defended on grounds of parsimony (its real defense is more likely to be analytic tractability), it may actually exclude a number of important properties of development. For example, one important aspect of non-stationarity is that of incremental learning, whereby a system increases some resource over time with learning, and which is related to constructivist learning. Elman (1991) has demonstrated an important property of incremental learning in applications to language acquisition. Specifically, Elman (1991) found that networks that start small by limiting some resource, such as working memory, could learn the structure of embedded sentences with long-distance dependencies, whereas larger networks could not. Briefly, an initially restricted network was more sensitive to the low-order statistics in the input data than was a larger network, as these resource limitations effectively acted as a filter on the input to simplify the initial input to the system. However, as these resources increased, the prior learning served as a constraint on subsequent learning and the network could effectively use the lower-order statistics it had learned as a basis to learn the higher-order statistics.[5]

This sort of learning clearly violates both the continuity hypothesis and the related instantaneous learning idealization of Chomsky (1965), but it helps to resolve a paradox that surrounds initial resource limitations and learnability. While it has been held that initial resource limitations actually make the problem of learning more difficult since they weaken the learning capacity of a system (e.g., Wexler & Culicover, 1980), Elman's demonstration illustrates how these initial limitations may facilitate learning. It also supports Newport's (1990) interpretation of her experimental findings that children are more efficient language learners because of these initial limitations. This, then, illustrates an important point regarding non-stationarity: a system that does not initially fully express its computational resources may learn in bootstrap fashion by using its initially limited resources to learn a subset of some domain that then constrains subsequent learning. Were that system to start by fully expressing its resources, it would not successfully learn the problem confronting it, but may simply fail by overfitting the data, a problem common to large networks (see Fahlman &

[5] Elman (1991) trained recurrent networks in which the feedback was initially restricted and then increased over training. This had the effect of a sliding temporal window over the network's access to its own states.

Lebiere, 1990), for a discussion of incremental learning in the context of constructivist networks).

7. Some learning-theoretic properties of constructivist models

The previous section illustrated that systems that incorporate non-stationarity may have important learning characteristics and that the charge that such systems trivialized explanations by introducing arbitrary changes was unfounded. Although I am most concerned only with demonstrating the plausibility of constructivism, in this section I briefly consider some of the learning-theoretic properties of constructivist models and contrast them with standard PDP models (see Quartz, 1993a, for more detail).

The representational and learning-theoretic properties of constructivist networks fundamentally differ from fixed networks. Perhaps most importantly, PDP networks are limited to an *a priori* defined, fixed hypothesis space, and are, therefore, model-based estimators that in complex domains may perform quite poorly. While this restriction results in fast learning in cases where an appropriate hypothesis space is chosen, it also leads to error in cases where the target function, or an acceptable approximation, is not contained in the hypothesis space. However, as Baum (1989, p. 203) points out, the fact that some concept *F* is not learnable relative to some representation class *G* only indicates that the wrong representation class may have been chosen, and that "a pragmatic learner should be willing to use any class of representations necessary to solve the problem".

This suggests that it would be advantageous for a system not to be limited to its *a priori* knowledge of some domain. Although constructivist networks have built into their initial state some hypothesis space, they are not limited to this representation class. Rather, since constructivist networks may build their architecture, and therefore representations, as they learn, they have been shown (Baum, 1988, 1989) to be "complete" representations, capable of learning any concept learnable in polynomial time by any representation that is computable in polynomial time, and which escape the NP-completeness results that afflict fixed architectures (Blum & Rivest, 1988; Judd, 1988) These NP-completeness results bring into serious doubt the ability of fixed networks to learn in polynomial time for large problems.

The demonstration that constructivist networks are "complete" representations suggests that the relativization of learning to some particular choice of representation to ensure feasibility may be relaxed for the constructivist learner to allow learning to be relative to the class of *all* polynomial computable representations. Essentially this maintains, then, that if *F* is learnable by any representation that can be computed in polynomial time, then F is learnable by a constructivist network.[6] Hence, by not being limited to some *a priori* defined, fixed hypothesis space, constructivist models can build representations as they learn to come to represent any concept learnable in polynomial time. In this sense, constructivist models are more general learners in contrast to PDP learners that relativize learning to some *a priori* fixed representation class.

Although these results point to the theoretical capacities of constructivist networks, they do not demonstrate their utility in applications. While constructivist networks have not been as extensively studied as standard PDP models, in benchmark tests constructivist models significantly outperform standard PDP back-propagation networks. For example, Fahlman and Lebiere (1990) examined the performance of their constructive cascade-correlation algorithm against standard back-propagation on the two-spirals problem, which has as its goal to classify training points of two interlocking spirals. They found that their network outperformed back-propagation by at least a factor of 10 and also outperformed back-propagation on the parity problem. In addition to outperforming PDP models, the cascade-correlation model automatically found efficient network topologies and showed incremental learning features. Although further comparisons are required, and benchmarks are complicated by the possibility of idiosyncratic or non-representative problems, these results suggest that constructivist algorithms have a number of performance advantages over standard PDP models.

8. A measure of representational power

One of the obstacles to establishing the plausibility of constructivism was the lack of an appropriate quantitative measure of a system's representational complexity that could be reasonably seen as increasing with learning. As I mentioned above, since neural networks identify representational properties with structural ones, this relation may suggest such a measure. In fact, approximation studies of the capacities of neural networks establish a direct relation between the ability of a network to approximate some function or concept and the complexity of the network that provides such a measure, as I consider below.

PDP networks are fundamentally limited in that they are model-based estimators that may only partially approximate some concept. In contrast, networks with the ability to add hidden units as a function of learning can learn arbitrarily accurate representations and thus are universal approximators (Hor-

[6] The condition that the representation be computable in polynomial time is an additional complexity-based demand that not only is the sample complexity polynomial but the time complexity – the time it takes the system to process the samples to arrive at a representation – also scales polynomially.

nik, Stinchcombe, & White, 1989). The upshot of these approximation studies is that networks with the ability to add units at an appropriate rate relative to the size of the training set n are capable of consistent (hence learnable) non-parametric regression (White, 1990). Although the details are not relevant here, in these approximation theory studies, the representational complexity of a network is indexed simply by the number of hidden units, q, and, allowing q to grow at an appropriate rate relative to the training set size n, it can be shown that such a network increases its representational complexity so that it may arbitrarily accurately approximate some function. As n increases, the network adds hidden units to build successively more powerful concepts until it converges on the target concept.

Hence, we may simply take q, the number of hidden units (or the number of units that perform the salient part of the computation), as an expedient measure of the representational power of a network that can replace Fodor's choice of logic as this measure, as I consider in more detail below.[7]

9. Conclusions

As the Valiant model illustrates, and as statistical studies have confirmed (Geman et al. 1992), the central problem of learning is perhaps not so much learning in the sense of statistical inference, but the more fundamental problem confronting a cognitive system is that of constructing appropriate representations to serve as the basis for the acquisition of those skills that define the mature state. The main promise of constructivism is that it uniquely allows for the structure of the learning system's environment to play a central role in the construction of the representations that underlie the system's ability to learn in that environment. The learning-theoretic positions outlined here suggest that this is a maximally powerful strategy that escapes the shortcomings of attempting to define these representations *a priori*. Indeed, the prolonged extent of postnatal human development, with its corresponding progressive increase in neural complexity, suggests that it is a highly adaptive strategy to allow environmental factors to directly influence brain structure, as 30 years of neurobiological research has suggested (reviewed in Greenough & Bailey, 1988; Juraska, 1990).

To return to Fodor's example of representational power, which I stated earlier was an important element of his position, it is true that a weaker logic cannot represent the primitives of a stronger logic; hence, there must be principled discontinuities between such systems. It is these principled discontinuities inher-

[7] From within the PAC framework, a measure of representational power of a class of concepts (or architecture) is the Vapnik–Chervonenkis (VC) dimension (see Abu-Mostafa, 1989). I consider only the number of hidden units q here to avoid introducing more technical definitions into the discussion and since q is a rough bound of the VC dimension.

ent in Fodor's choice of logic as the quantitative measure of a conceptual structure's representational power that appears to make the problem of concept acquisition intractable. However, once we abandon the notion that the expressive power of a structure must be construed in terms of such a logic, there are natural ways to show that more powerful structures may be acquired on the basis of simpler ones. For example, by allowing a network to add new connections and units as a function of learning, as many constructivist algorithms have explored (Fahlman, 1991; Fahlman & Lebiere, 1990; Frean, 1990; Gallant, 1986; Nadel, 1989; Wynne-Jones, 1993), such a network may extend its representation class beyond its initial state to include novel concepts. And, since there is an immediate relation between increases in the complexity of a neural network architecture and its representational power, indexing representational power by the number of hidden units, q (or the number of units that perform the salient part of the computation), provides a natural measure to see how increases in the complexity of the architecture lead to increases in this measure of representational power. Although constructivist learning is not precluded in classical or symbolic architectures, it is this immediate relation between structure and function in neural networks that makes the plausibility of constructivist networks evident by showing how structural increases may lead to novel representations.

A number of empirical issues raised by this discussion remain untouched (Quartz, 1993b). However, my main aim in this paper has only been to demonstrate the plausibility, or coherence, of the constructivist position. Fodor's argument is a theoretical one, aimed not at an appraisal of the empirical support for various models of acquisition, but only at the logical coherence of the positions themselves. Thus, the interpretation of constructivism in terms of neural networks with the ability to add new connections and units as a function of learning represents both a natural interpretation of constructivism in terms of possible mechanisms and a logically coherent one. In addition, the theoretical results I briefly discussed suggest that constructivist networks are powerful learners in contrast to the fundamental limitations of PDP models, which were shown to be nativist. However, constructivist class of learners has not been extensively explored, in part because of the widely held assumption that they were not plausible models. Perhaps with their natural interpretation in terms of neural network learning, constructivist models will come to play a more prominent role in neural network research and in cognitive science.

References

Abu-Mostafa, Y.S. (1989). The Vapnik–Chervonenkis dimension: information versus complexity in learning. *Neural Computation*, *1*, 312–317.

Bates, E. (1992). Language development. *Current Opinion in Neurobiology*, *2*, 180–185.

Baum, E.B. (1988). Complete representations for learning from examples. In Y. Abu-Mostafa (Ed.), *Complexity in information theory*, (pp. 77–98). Heidelberg: Springer-Verlag.

Baum, E.B. (1989). A proposal for more powerful learning algorithms. *Neural Computation*, *1*, 201–207.

Baum, E.B., & Haussler, D. (1989). What size net gives valid generalization? *Neural Computation*, *1*, 151–160.

Black, J.E., Isaacs, K.R., Anderson B.J., Alcantara, A.A., & Greenough, W.T. (1990). Learning causes synaptogenesis, whereas motor activity causes angiogenesis, in cerebellar cortex of adult rats. *Proceedings of the National Academy of Sciences, USA*, *87*, 5568–5572.

Blum, A., & Rivest, R.L. (1988). Training a 3-node neural network is NP-complete. In *Proceedings of the 1988 Workshop on Computational Learning Theory* (pp. 9–18). San Mateo, CA: Morgan-Kaufmann.

Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. (1987). Learnability and the Vapnik–Chervonenkis dimension. *UCSC Technical Report*, UCSC-CRL-87-20.

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

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, MA: MIT Press.

Dente, J., & Mendes, R.V. (1992). Learning from examples and generalization. *Complex Systems*, 6, 301–314.

Dietterich, T.G. (1990). Machine learning. *Annual Review of Computer Science*, *4*, 255–306.

Elman, J. (1991). Incremental learning, or the importance of starting small. In *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society* (pp. 443–448). Hillsdale, NJ: Erlbaum.

Fahlman, S.E. (1991). The recurrent cascade-correlation architecture. In R.P. Lippmann & J.E. Moody (Eds.), *Advances in neural information processing systems 3*. San Mateo: Morgan Kauffmann.

Fahlmann, S.E. & Lebiere, C. (1990). The cascade-correlation learning architecture. In D.S. Touretzky (Ed.), *Advances in neural information processing systems 2*. San Mateo: Morgan Kauffmann.

Fodor, J. (1975). *The language of thought*. Cambridge, MA: Harvard University Press.

Fodor, J. (1980). Fixation of belief and concept acquisition. In M. Piattelli-Palmarini (Ed.), *Language and learning: the debate between Chomsky and Piaget* (pp. 143–149). Cambridge, MA: Harvard Press.

Fodor, J. (1981). The current status of the innateness controversy. In *Representations* (pp. 257–316). Cambridge, MA: MIT Press.

Frean, M. (1990). The upstart algorithm: a method for constructing and training feedforward neural networks. *Neural Computation*, *2*, 198–209.

Gallant, S.I. (1986). Three constructive algorithms for neural network learning. In *Annual Meeting of the Cognitive Science Society* (pp. 652–660). Hillsdale, NJ: Erlbaum.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, 4, 1–58.

Greenough, W.T., & Bailey, C. (1988). The anatomy of memory: convergence of results across a diversity of tests. *Trends in Neurosciences*, *11*, 142–147.

Gold, E.M. (1967). Language identification in the limit. *Information and Control*, *10*, 447–474.

Hebb, D.O. (1949). *The organization of behavior: a neuropsychological theory*. New York: Wiley.

Hertz, J., Krogh, A., & Palmer, R.G. (1991). *Introduction to the theory of neural computation*. Redwood City, CA: Addison-Wesley.

Hockfield, S., & Kalb, R.G. (1993). Activity-dependent structural changes during neuronal developments. *Current Opinion in Neurobiology*, *3*, 87–92.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*, 359–366.

Judd, S. (1988). On the complexity of loading shallow neural networks. *Journal of Complexity*, *4*, 177–192.

Juraska, J.M. (1990). The structure of the rat cerebral cortex: effects of gender and environment. In

B. Kolb & R.C. Tees (Eds.), *The cerebral cortex of the rat*, pp. 483–506. Cambridge, MA: MIT Press.

Lightfoot, D. (1991). The child's trigger experience: degree-0 learnability. *Behavioral and Brain Sciences*, *14*, 364–394.

Lightfoot, D. (1992). *How to set parameters*. Cambridge, MA: MIT Press.

Macnamara, J. (1982). *Names for things: a study of child language*. Cambridge, MA: MIT Press.

Nadel, J.P. (1989). Study of a growth algorithm for a feed forward network. *International Journal of Neural Systems*, *1*, 55–59.

Natarajan, B. (1991). *Machine learning: a theoretical approach*. San Mateo, CA: Morgan Kaufmann.

Newport, E.L. (1990). Maturational constraints on language learning. *Cognitive Science*, *14*, 11–28.

Piaget, J. (1954). *The construction of reality in the child* (M. Cook, Trans.). New York: Basic Books.

Piattelli-Palmarini, M. (Ed.) (1980). *Language and learning: the debate between Jean Piaget and Noam Chomsky*. Cambridge, UK: Harvard University Press.

Piattelli-Palmarini, M. (1989). Evolution, selection and cognition: from "learning" to parameter setting in biology and in the study of language. *Cognition*, *31*, 1–44.

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

Pinker, S. (1981). Comments on paper by Wexler. In C.L. Baker & J.J. McCarthy (Eds.), *The logical problem of language acquisition*. Cambridge, MA: MIT Press.

Pinker, S. (1984). *Language learnability and language development*. Cambridge, MA: Harvard University Press.

Pinker S., & Prince, A. (1988). On language and connectionism: an analysis of a parallel distributed processing model of language acquisition. *Cognition*, *28*, 73–193.

Pylyshyn, Z. (1984). *Computation and cognition*. Cambridge, MA: MIT Press.

Quartz, S.R. (1993a). Learnability, neural networks, and the problem of development. Submitted.

Quartz, S.R. (1993b). A constructivist model of brain development: systems, cellular, and molecular evidence. Submitted.

Rumelhart, D.E., & McClelland, J.L. (1986). *Parallel distributed processing: explorations in the microstructure of cognition*. Cambridge, MA: MIT Press.

Sejnowski, T.J., & Tesauro, G. (1989). The Hebb rule for synaptic plasticity: algorithms and implementations. In J.H. Byrne & W.O. Berry (Eds.) *Neural models of plasticity* (pp. 94–103). San Diego, CA: Academic Press.

Shatz, C.J. (1990). Impulse activity and patterning of connections during development. *Neuron*, *5*, 745–756.

Shepherd, G.M. (1990). The significance of real neuron architectures for neural network simulations. In E. Schwartz (Ed.), *Computational neuroscience* (pp. 82–96). Cambridge, MA: MIT Press.

Valiant, L.G. (1984). A theory of the learnable. *Communications of the ACM*, *27*, 1134–1142.

Wexler, K., & Culicover, P. (1980). *Formal principles of language acquisition*. Cambridge, MA: MIT Press.

White, H. (1990). Connectionist nonparametric regression: multilayer feedforward networks can learn arbitrary mappings. *Neural Networks*, *3*, 535–549.

Wynne-Jones, M. (1993). Node splitting: a constructive algorithm for feed-forward neural networks. *Neural Computing and Applications*, *1*, 17–22.