Connectionism

First published Sun May 18, 1997; substantive revision Thu Feb 19, 2015

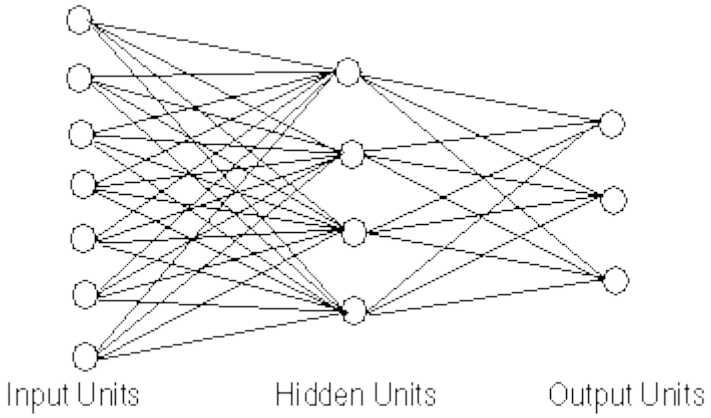
Connectionism is a movement in cognitive science that hopes to explain intellectual abilities using artificial neural networks (also known as 'neural networks' or 'neural nets'). Neural networks are simplified models of the brain composed of large numbers of units (the analogs of neurons) together with weights that measure the strength of connections between the units. These weights model the effects of the synapses that link one neuron to another. Experiments on models of this kind have demonstrated an ability to learn such skills as face recognition, reading, and the detection of simple grammatical structure.

Philosophers have become interested in connectionism because it promises to provide an alternative to the classical theory of the mind: the widely held view that the mind is something akin to a digital computer processing a symbolic language. Exactly how and to what extent the connectionist paradigm constitutes a challenge to classicism has been a matter of hot debate in recent years.

1. A Description of Neural Networks

A neural network consists of large number of units joined together in a pattern of connections. Units in a net are usually segregated into three classes: input units, which receive information to be processed, output units where the results of the processing are found, and units in between called hidden units. If a neural net were to model the whole human nervous system, the input units would be analogous to the sensory neurons, the output units to the motor neurons, and the hidden units to all other neurons.

Here is a simple illustration of a simple neural net:



Each input unit has an activation value that represents some feature external to the net. An input unit sends its activation value to each of the hidden units to which it is connected. Each of these hidden units calculates its own activation value depending on the activation values it receives from the input units. This signal is then passed on to output units or to another layer of hidden units. Those hidden units compute their activation values in the same way, and send them along to their neighbors. Eventually the signal at the input units propagates all the way through the net to determine the activation values at all the output units.

The pattern of activation set up by a net is determined by the weights, or strength of connections between the units. Weights may be either positive or negative. A negative weight represents the

inhibition of the receiving unit by the activity of a sending unit. The activation value for each receiving unit is calculated according a simple activation function. Activation functions vary in detail, but they all conform to the same basic plan. The function sums together the contributions of all sending units, where the contribution of a unit is defined as the weight of the connection between the sending and receiving units times the sending unit's activation value. This sum is usually modified further, for example, by adjusting the activation sum to a value between 0 and 1 and/or by setting the activation to zero unless a threshold level for the sum is reached. Connectionists presume that cognitive functioning can be explained by collections of units that operate in this way. Since it is assumed that all the units calculate pretty much the same simple activation function, human intellectual accomplishments must depend primarily on the settings of the weights between the units.

The kind of net illustrated above is called a feed forward net. Activation flows directly from inputs to hidden units and then on to the output units. More realistic models of the brain would include many layers of hidden units, and recurrent connections that send signals back from higher to lower levels. Such recurrence is necessary in order to explain such cognitive features as short-term memory. In a feed forward net, repeated presentations of the same input produce the same output every time, but even the simplest organisms habituate to (or learn to ignore) repeated presentation of the same stimulus. Connectionists tend to avoid recurrent connections because little is understood about the general problem of training recurrent nets. However Elman (1991) and others have made some progress with simple recurrent nets, where the recurrence is tightly constrained.

2. Neural Network Learning and Backpropagation

Finding the right set of weights to accomplish a given task is the central goal in connectionist research. Luckily, learning algorithms have been devised that can calculate the right weights for carrying out many tasks. (See Hinton 1992 for an accessible review.) These fall into two broad categories: supervised and unsupervised learning. Hebbian learning is the best known unsupervised form.

As each input is presented to the net, weights between nodes that are active together are increased, while those weights connecting nodes that are not active together are decreased. This form of training is especially useful for building nets that can classify the input into useful categories. The most widely used supervised algorithm is called backpropagation. To use this method, one needs a training set consisting of many examples of inputs and their desired outputs for a given task. This external set of examples "supervises" the training process. One of the most widely used of these training methods is called backpropagation. To use this method one needs a training set consisting of many examples of inputs and their desired outputs for a given task. If, for example, the task is to distinguish male from female faces, the training set might contain pictures of faces together with an indication of the sex of the person depicted in each one. A net that can learn this task might have two output units (indicating the categories male and female) and many input units, one devoted to the brightness of each pixel (tiny area) in the picture. The weights of the net to be trained are initially set to random values, and then members of the training set are repeatedly exposed to the net. The values for the input of a member are placed on the input units and the output of the net is compared with the desired output for this member. Then all the weights in the net are adjusted slightly in the direction that would bring the net's output values closer to the values for the desired output. For example, when male's face is presented to the input units the weights are adjusted so that the value of the male output unit is increased and the value of the female output unit is decreased. After many repetitions of this process the net may learn to produce the desired output for each input in the training set. If the training goes well, the net may also have learned to generalize to the desired behavior for inputs and outputs that were not in the training set. For example, it may do a good job of distinguishing males from females in pictures that were never presented to it before.

Training nets to model aspects of human intelligence is a fine art. Success with backpropagation and other connectionist learning methods may depend on quite subtle adjustment of the algorithm and the training set. Training typically involves hundreds of thousands of rounds of weight adjustment. Given the limitations of computers presently available to connectionist researchers, training a net to perform an interesting task may take days or even weeks. Some of the difficulty may be resolved when parallel circuits

specifically designed to run neural network models are widely available. But even here, some limitations to connectionist theories of learning will remain to be faced. Humans (and many less intelligent animals) display an ability to learn from single events; for example an animal that eats a food that later causes gastric distress will never try that food again. Connectionist learning techniques such as backpropagation are far from explaining this kind of 'one shot' learning.

3. Samples of What Neural Networks Can Do

Connectionists have made significant progress in demonstrating the power of neural networks to master cognitive tasks. Here are three well-known experiments that have encouraged connectionists to believe that neural networks are good models of human intelligence. One of the most attractive of these efforts is Sejnowski and Rosenberg's 1987 work on a net that can read English text called NET talk. The training set for NET talk was a large data base consisting of English text coupled with its corresponding phonetic output, written in a code suitable for use with a speech synthesizer. Tapes of NET talk's performance at different stages of its training are very interesting listening. At first the output is random noise. Later, the net sounds like it is babbling, and later still as though it is speaking English double-talk (speech that is formed of sounds that resemble English words). At the end of training, NET talk does a fairly good job of pronouncing the text given to it. Furthermore, this ability generalizes fairly well to text that was not presented in the training set.

Another influential early connectionist model was a net trained by Rumelhart and McClelland (1986) to predict the past tense of English verbs. The task is interesting because although most of the verbs in English (the regular verbs) form the past tense by adding the suffix '-ed', many of the most frequently verbs are irregular ('is' / 'was', 'come' / 'came', 'go' / 'went'). The net was first trained on a set containing a large number of irregular verbs, and later on a set of 460 verbs containing mostly regulars. The net learned the past tenses of the 460 verbs in about 200 rounds of training, and it generalized fairly well to verbs not in the training set. It even showed a good appreciation of "regularities" to be found among the irregular verbs ('send' / 'sent', 'build' / 'built'; 'blow' /

‘blew’, ‘fly’ / ‘flew’). During learning, as the system was exposed to the training set containing more regular verbs, it had a tendency to overregularize, i.e., to combine both irregular and regular forms: (‘break’ / ‘broked’, instead of ‘break’ / ‘broke’). This was corrected with more training. It is interesting to note that children are known to exhibit the same tendency to overregularize during language learning. However, there is hot debate over whether Rumelhart and McClelland's is a good model of how humans actually learn and process verb endings. For example, Pinker & Prince (1988) point out that the model does a poor job of generalizing to some novel regular verbs. They believe that this is a sign of a basic failing in connectionist models. Nets may be good at making associations and matching patterns, but they have fundamental limitations in mastering general rules such as the formation of the regular past tense. These complaints raise an important issue for connectionist modelers, namely whether nets can generalize properly to master cognitive tasks involving rules. Despite Pinker and Prince's objections, many connectionists believe that generalization of the right kind is still possible (Niklasson and van Gelder 1994).

Elman's 1991 work on nets that can appreciate grammatical structure has important implications for the debate about whether neural networks can learn to master rules. Elman trained a simple recurrent network to predict the next word in a large corpus of English sentences. The sentences were formed from a simple vocabulary of 23 words using a subset of English grammar. The grammar, though simple, posed a hard test for linguistic awareness. It allowed unlimited formation of relative clauses while demanding agreement between the head noun and the verb. So for example, in the sentence

Any **man** that chases dogs that chase cats … run**s**.

the singular ‘**man**’ must agree with the verb ‘run**s**’ despite the intervening plural nouns (‘dogs’, ‘cats’) which might cause the selection of ‘run’. One of the important features of Elman's model is the use of recurrent connections. The values at the hidden units are saved in a set of so called context units, to be sent back to the input level for the next round of processing. This looping back from hidden to input layers provides the net with a rudimentary form of memory of the sequence of words in the input sentence. Elman's nets displayed an appreciation of the grammatical structure of sentences that were not in the training set. The net's command of syntax was measured in the following way. Predicting the next

word in an English sentence is, of course, an impossible task. However, these nets succeeded, at least by the following measure. At a given point in an input sentence, the output units for words that are grammatical continuations of the sentence at that point should be active and output units for all other words should be inactive. After intensive training, Elman was able to produce nets that displayed perfect performance on this measure including sentences not in the training set. The work of Christiansen and Chater (1999a) and Morris et al. (2000) extends this research to more complex grammars. For a broader view of progress in connectionist natural language processing see summaries by Christiansen and Chater (1999b), and Rhode and Plaut (2004).

Although this performance is impressive, there is still a long way to go in training nets that can process a language like English. Furthermore, doubts have been raised about the significance of Elman's results. For example, Marcus (1998, 2001) argues that Elman's nets are not able to generalize this performance to sentences formed from a novel vocabulary. This, he claims, is a sign that connectionist models merely associate instances, and are unable to truly master abstract rules. On the other hand, Phillips (2002) argues that classical architectures are no better off in this respect. The purported inability of connectionist models to generalize performance in this way has become an important theme in the systematicity debate. (See Section 7 below.)

A somewhat different concern about the adequacy of connectionist language processing focuses on tasks that mimic infant learning of simple artificial grammars. Data on reaction time confirms that infants can learn to distinguish well-formed from ill-formed sentences in a novel language created by experimenters. Shultz and Bale (2001) report success in training neural nets on the same task. Vilcu and Hadley (2005) object that this work fails to demonstrate true acquisition of the grammar, but see Shultz and Bale (2006) for a detailed reply.

4. Strengths and Weaknesses of Neural Network Models

Philosophers are interested in neural networks because they may provide a new framework for understanding the nature of the mind and its relation to the brain (Rumelhart and McClelland 1986,

Chapter 1). Connectionist models seem particularly well matched to what we know about neurology. The brain is indeed a neural net, formed from massively many units (neurons) and their connections (synapses). Furthermore, several properties of neural network models suggest that connectionism may offer an especially faithful picture of the nature of cognitive processing. Neural networks exhibit robust flexibility in the face of the challenges posed by the real world. Noisy input or destruction of units causes graceful degradation of function. The net's response is still appropriate, though somewhat less accurate. In contrast, noise and loss of circuitry in classical computers typically result in catastrophic failure. Neural networks are also particularly well adapted for problems that require the resolution of many conflicting constraints in parallel. There is ample evidence from research in artificial intelligence that cognitive tasks such as object recognition, planning, and even coordinated motion present problems of this kind. Although classical systems are capable of multiple constraint satisfaction, connectionists argue that neural network models provide much more natural mechanisms for dealing with such problems.

Over the centuries, philosophers have struggled to understand how our concepts are defined. It is now widely acknowledged that trying to characterize ordinary notions with necessary and sufficient conditions is doomed to failure. Exceptions to almost any proposed definition are always waiting in the wings. For example, one might propose that a tiger is a large black and orange feline. But then what about albino tigers? Philosophers and cognitive psychologists have argued that categories are delimited in more flexible ways, for example via a notion of family resemblance or similarity to a prototype. Connectionist models seem especially well suited to accommodating graded notions of category membership of this kind. Nets can learn to appreciate subtle statistical patterns that would be very hard to express as hard and fast rules. Connectionism promises to explain flexibility and insight found in human intelligence using methods that cannot be easily expressed in the form of exception free principles (Horgan and Tienson 1989, 1990), thus avoiding the brittleness that arises from standard forms of symbolic representation.

Despite these intriguing features, there are some weaknesses in connectionist models that bear mentioning. First, most neural network research abstracts away from many interesting and

possibly important features of the brain. For example, connectionists usually do not attempt to explicitly model the variety of different kinds of brain neurons, nor the effects of neurotransmitters and hormones. Furthermore, it is far from clear that the brain contains the kind of reverse connections that would be needed if the brain were to learn by a process like backpropagation, and the immense number of repetitions needed for such training methods seems far from realistic. Attention to these matters will probably be necessary if convincing connectionist models of human cognitive processing are to be constructed. A more serious objection must also be met. It is widely felt, especially among classicists, that neural networks are not particularly good at the kind of rule based processing that is thought to undergird language, reasoning, and higher forms of thought. (For a well known critique of this kind see Pinker and Prince 1988.) We will discuss the matter further when we turn to the systematicity debate.

5. The Shape of the Controversy between Connectionists and Classicists

The last forty years have been dominated by the classical view that (at least higher) human cognition is analogous to symbolic computation in digital computers. On the classical account, information is represented by strings of symbols, just as we represent data in computer memory or on pieces of paper. The connectionist claims, on the other hand, that information is stored non-symbolically in the weights, or connection strengths, between the units of a neural net. The classicist believes that cognition resembles digital processing, where strings are produced in sequence according to the instructions of a (symbolic) program. The connectionist views mental processing as the dynamic and graded evolution of activity in a neural net, each unit's activation depending on the connection strengths and activity of its neighbors, according to the activation function.

On the face of it, these views seem very different. However many connectionists do not view their work as a challenge to classicism and some overtly support the classical picture. So-called implementational connectionists seek an accommodation between the two paradigms. They hold that the brain's net implements a

symbolic processor. True, the mind is a neural net; but it is also a symbolic processor at a higher and more abstract level of description. So the role for connectionist research according to the implementationalist is to discover how the machinery needed for symbolic processing can be forged from neural network materials, so that classical processing can be reduced to the neural network account.

However, many connectionists resist the implementational point of view. Such radical connectionists claim that symbolic processing was a bad guess about how the mind works. They complain that classical theory does a poor job of explaining graceful degradation of function, holistic representation of data, spontaneous generalization, appreciation of context, and many other features of human intelligence which are captured in their models. The failure of classical programming to match the flexibility and efficiency of human cognition is by their lights a symptom of the need for a new paradigm in cognitive science. So radical connectionists would eliminate symbolic processing from cognitive science forever.

The controversy between radical and implementational connectionists is complicated by the invention of what are called hybrid connectionist architectures. Here elements of classical symbolic processing are included in neural nets (Wermter and Sun, 2000). For example, Miikkulainen (1993) champions a complex collection of neural net modules that share data coded in activation patterns. Since one of the modules acts as a memory, the system taken as a whole resembles a classical processor with separate mechanisms for storing and operating on digital “words”. Smolensky (1991) is famous for inventing so called tensor product methods for simulating the process of variable binding, where symbolic information is stored at and retrieved from known “locations”. More recently, Eliasmith (2013) has proposed complex and massive architectures that use what are called semantic pointers, which exhibit features of classical variable binding. Once hybrid architectures such as these are on the table, it becomes more difficult to classify a given connectionist model as radical or merely implementational. This opens the interesting prospect that whether symbolic processing is actually present in the human brain may turn out to be a matter of degree.

6. Connectionist Representation

Connectionist models provide a new paradigm for understanding how information might be represented in the brain. A seductive but naive idea is that single neurons (or tiny neural bundles) might be devoted to the representation of each thing the brain needs to record. For example, we may imagine that there is a grandmother neuron that fires when we think about our grandmother. However, such local representation is not likely. There is good evidence that our grandmother thought involves complex patterns of activity distributed across relatively large parts of cortex.

It is interesting to note that distributed, rather than local representations on the hidden units are the natural products of connectionist training methods. The activation patterns that appear on the hidden units while NETtalk processes text serve as an example. Analysis reveals that the net learned to represent such categories as consonants and vowels, not by creating one unit active for consonants and another for vowels, but rather in developing two different characteristic patterns of activity across all the hidden units.

Given the expectations formed from our experience with local representation on the printed page, distributed representation seems both novel and difficult to understand. But the technique exhibits important advantages. For example, distributed representations, (unlike symbols stored in separate fixed memory locations) remain relatively well preserved when parts of the model are destroyed or overloaded. More importantly, since representations are coded in patterns rather than firings of individual units, relationships between representations are coded in the similarities and differences between these patterns. So the internal properties of the representation carry information on what it is about (Clark 1993, 19). In contrast, local representation is conventional. No intrinsic properties of the representation (a unit's firing) determine its relationships to the other symbols. This self-reporting feature of distributed representations promises to resolve a philosophical conundrum about meaning. In a symbolic representational scheme, all representations are composed out of symbolic atoms (like words in a language). Meanings of complex symbol strings may be defined by the way they are built up out of their constituents, but what fixes the meanings of the atoms?

Connectionist representational schemes provide an end run around the puzzle by simply dispensing with atoms. Every distributed representation is a pattern of activity across all the units, so there is no principled way to distinguish between simple and complex representations. To be sure, representations are composed out of the activities of the individual units. But none of these 'atoms' codes for any symbol. The representations are sub-symbolic in the sense that analysis into their components leaves the symbolic level behind.

The sub-symbolic nature of distributed representation provides a novel way to conceive of information processing in the brain. If we model the activity of each neuron with a number, then the activity of the whole brain can be given by a giant vector (or list) of numbers, one for each neuron. Both the brain's input from sensory systems and its output to individual muscle neurons can also be treated as vectors of the same kind. So the brain amounts to a vector processor, and the problem of psychology is transformed into questions about which operations on vectors account for the different aspects of human cognition.

Sub-symbolic representation has interesting implications for the classical hypothesis that the brain must contain symbolic representations that are similar to sentences of a language. This idea, often referred to as the language of thought (or LOT) thesis may be challenged by the nature of connectionist representations. It is not easy to say exactly what the LOT thesis amounts to, but van Gelder (1990) offers an influential and widely accepted benchmark for determining when the brain should be said to contain sentence-like representations. It is that when a representation is tokened one thereby tokens the constituents of that representation. For example, if I write 'John loves Mary' I have thereby written the sentence's constituents: 'John' 'loves' and 'Mary'. Distributed representations for complex expressions like 'John loves Mary' can be constructed that do not contain any explicit representation of their parts (Smolensky 1991). The information about the constituents can be extracted from the representations, but neural network models do not need to explicitly extract this information themselves in order to process it correctly (Chalmers 1990). This suggests that neural network models serve as counterexamples to the idea that the language of thought is a prerequisite for human cognition. However, the matter is still a topic of lively debate (Fodor 1997).

The novelty of distributed and superimposed connectionist information storage naturally causes one to wonder about the viability of classical notions of symbolic computation in describing the brain. Ramsey (1997) argues that though we may attribute symbolic representations to neural nets, those attributions do not figure in legitimate explanations of the model's behavior. This claim is important because the classical account of cognitive processing, (and folk intuitions) presume that representations play an explanatory role in understanding the mind. It has been widely thought that cognitive science requires, by its very nature, explanations that appeal to representations (Von Eckardt 2003). If Ramsey is right, the point may cut in two different ways. Some may use it to argue for a new and non-classical understanding of the mind, while others would use it to argue that connectionism is inadequate since it cannot explain what it must. However, Haybron (2000) argues against Ramsey that there is ample room for representations with explanatory role in radical connectionist architectures. Roth (2005) makes the interesting point that contrary to first impressions, it may also make perfect sense to explain a net's behavior by reference to a computer program, even if there is no way to discriminate a sequence of steps of the computation through time.

The debate concerning the presence of classical representations and a language of thought has been clouded by lack of clarity in defining what should count as the representational “vehicles” in distributed neural models. Shea (2007) makes the point that the individuation of distributed representations should be defined by the way activation patterns on the hidden units cluster together. It is the relationships between clustering *regions* in the space of possible activation patterns that carry representational content, not the activations themselves, nor the collection of units responsible for the activation. On this understanding, prospects are improved for locating representational content in neural nets that can be compared in nets of different architectures, that is causally involved in processing, and which overcomes some objections to holistic accounts of meaning.

In a series of papers Horgan and Tienson (1989, 1990) have championed a view called representations without rules. According to this view classicists are right to think that human brains (and good connectionist models of them) contain explanatorily robust representations; but they are wrong to think that those

representations enter in to hard and fast rules like the steps of a computer program. The idea that connectionist systems may follow graded or approximate regularities (“soft laws” as Horgan and Tienson call them) is intuitive and appealing. However, Aizawa (1994) argues that given an arbitrary neural net with a representation level description, it is always possible to outfit it with hard and fast representation-level rules. Guarini (2001) responds that if we pay attention to notions of rule following that are useful to cognitive modeling, Aizawa's constructions will seem beside the point.

7. The Systematicity Debate

The major points of controversy in the philosophical literature on connectionism have to do with whether connectionists provide a viable and novel paradigm for understanding the mind. One complaint is that connectionist models are only good at processing associations. But such tasks as language and reasoning cannot be accomplished by associative methods alone and so connectionists are unlikely to match the performance of classical models at explaining these higher-level cognitive abilities. However, it is a simple matter to prove that neural networks can do anything that symbolic processors can do, since nets can be constructed that mimic a computer's circuits. So the objection can not be that connectionist models are unable to account for higher cognition; it is rather that they can do so only if they implement the classicist's symbolic processing tools. Implementational connectionism may succeed, but radical connectionists will never be able to account for the mind.

Fodor and Pylyshyn's often cited paper (1988) launches a debate of this kind. They identify a feature of human intelligence called systematicity which they feel connectionists cannot explain. The systematicity of language refers to the fact that the ability to produce/understand/think some sentences is intrinsically connected to the ability to produce/understand/think others of related structure. For example, no one with a command of English who understands ‘John loves Mary’ can fail to understand ‘Mary loves John.’ From the classical point of view, the connection between these two abilities can easily be explained by assuming that masters of English represent the constituents (‘John’, ‘loves’ and ‘Mary’) of ‘John loves Mary’ and compute its meaning from the

meanings of these constituents. If this is so, then understanding a novel sentence like 'Mary loves John' can be accounted for as another instance of the same symbolic process. In a similar way, symbolic processing would account for the systematicity of reasoning, learning and thought. It would explain why there are no people who are capable of concluding *P* from *P* & (*Q* & *R*), but incapable of concluding *P* from *P* & *Q*, why there are no people capable of learning to prefer a red cube to green square who cannot learn to prefer a green cube to the red square, and why there isn't anyone who can think that John loves Mary who can't also think that Mary loves John.

Fodor and McLaughlin (1990) argue in detail that connectionists do not account for systematicity. Although connectionist models can be trained to be systematic, they can also be trained, for example, to recognize 'John loves Mary' without being able to recognize 'Mary loves John.' Since connectionism does not guarantee systematicity, it does not explain why systematicity is found so pervasively in human cognition. Systematicity may exist in connectionist architectures, but where it exists, it is no more than a lucky accident. The classical solution is much better, because in classical models, pervasive systematicity comes for free.

The charge that connectionist nets are disadvantaged in explaining systematicity has generated a lot of interest. Chalmers (1993) points out that Fodor and Pylyshyn's argument proves too much, for it entails that all neural nets, even those that implement a classical architecture, do not exhibit systematicity. Given the uncontroversial conclusion that the brain is a neural net, it would follow that systematicity is impossible in human thought. Another often mentioned point of rebuttal (Aizawa 1997; Matthews 1997; Hadley 1997b) is that classical architectures do no better at explaining systematicity. There are also classical models that can be programmed to recognize 'John loves Mary' without being able to recognize 'Mary loves John,' for this depends on exactly which symbolic rules govern the classical processing. The point is that neither the use of connectionist architecture alone nor the use of classical architecture alone enforces a strong enough constraint to explain pervasive systematicity. In both architectures, further assumptions about the nature of the processing must be made to ensure that 'Mary loves John' and 'John loves Mary' are treated alike.

A discussion of this point should mention Fodor and McLaughlin's requirement that systematicity be explained as a matter of nomic necessity, that is, as a matter of natural law. The complaint against connectionists is that while they may implement systems that exhibit systematicity, they will not have explained it unless it follows from their models as a nomic necessity. However, the demand for nomic necessity is a very strong one, and one that classical architectures clearly cannot meet either. So the only tactic for securing a telling objection to connectionists along these lines would be to weaken the requirement on the explanation of systematicity to one which classical architectures can and connectionists cannot meet. A convincing case of this kind has yet to be made.

As the systematicity debate has evolved, attention has been focused on defining the benchmarks that would answer Fodor and Pylyshyn's challenge. Hadley (1994a, 1994b) distinguishes three brands of systematicity. Connectionists have clearly demonstrated the weakest of these by showing that neural nets can learn to correctly recognize novel sequences of words (e.g., 'Mary loves John') that were not in the training set. However, Hadley claims that a convincing rebuttal must demonstrate strong systematicity, or better, strong semantical systematicity. Strong systematicity would require (at least) that 'Mary loves John' be recognized even if 'Mary' never appears in the subject position in any sentence in the training set. Strong semantical systematicity would require as well that the net show abilities at correct semantical processing of the novel sentences rather than merely distinguishing grammatical from ungrammatical forms. Niklasson and van Gelder (1994) have claimed success at strong systematicity, though Hadley complains that this is at best a borderline case. Hadley and Hayward (1997) tackle strong semantical systematicity, but by Hadley's own admission it is not clear that they have avoided the use of a classical architecture. Boden and Niklasson (2000) claim to have constructed a model that meets at least the spirit of strong semantical systematicity, but Hadley (2004) argues that even strong systematicity has not been demonstrated there. Whether one takes a positive or a negative view of these attempts, it is safe to say that no one has met the challenge of providing a neural net capable of learning complex semantical processing that generalizes to a full range of truly novel inputs.

Research on nets that clearly demonstrate strong systematicity has continued. Jansen and Watter (2012) provide a good summary of more recent efforts along these lines, and propose an interesting basis for solving the problem. They use a more complex architecture that combines unsupervised self-organizing maps with features of simple recurrent nets. However, the main innovation is to allow codes for the words being processed to represent sensory-motor features of what the words represent. Once trained, their nets displayed very good accuracy in distinguishing the grammatical features of sentences whose words never even appeared in the training set. This may appear to be cheating since the word codes might surreptitiously represent grammatical categories, or at least they may unfairly facilitate learning those categories. Jansen and Watter note however, that the sensory-motor features of what a word represents are apparent to a child who has just acquired a new word, and so that information is not off-limits in a model of language learning. They make the interesting observation that a solution to the systematicity problem may require including sources of environmental information that have so far been ignored in theories of language learning. This work complicates the systematicity debate, since it opens a new worry about what information resources are legitimate in responding to the challenge. However, this reminds us that architecture alone (whether classical or connectionist) is not going to solve the systematicity problem in any case, so the interesting questions concern what sources of supplemental information are needed to make the learning of grammar possible.

Kent Johnson (2004) argues that the whole systematicity debate is misguided. Attempts at carefully defining the systematicity of language or thought leaves us with either trivialities or falsehoods. Connectionists surely have explaining to do, but Johnson recommends that it is fruitless to view their burden under the rubric of systematicity. Aizawa (2014) also suggests the debate is no longer germaine given the present climate in cognitive science. What is needed instead is the development of neurally plausible connectionist models capable of processing a language with a recursive syntax, which react immediately to the introduction of new items in the lexicon without introducing the features of classical architecture. The ‘systematicity’ debate may have already gone as Johnson advises, for Hadley’s demand for strong

semantical systematicity may be thought of as the requirement that connectionists exhibit success in that direction.

It has been over twenty-five years since the systematicity debate first began, with over 2,600 citations to Fodor and Pylyshyn's original paper. So this brief account is necessarily incomplete, and no doubt, biased. Aizawa (2003) provides an excellent view of the literature, and Calvo and Symons (2014) serves as another more recent resource.

8. Connectionism and Semantic Similarity

One of the attractions of distributed representations in connectionist models is that they suggest a solution to the problem of determining the meanings of brain states. The idea is that the similarities and differences between activation patterns along different dimensions of neural activity record semantical information. In this way, the similarity properties of neural activations provide intrinsic properties that fix meaning. However, Fodor and Lepore (1992, Ch. 6) challenge similarity based accounts on two fronts. The first problem is that human brains presumably vary significantly in the number of and connections between their neurons. Although it is straightforward to define similarity measures on two nets that contain the same number of units, it is harder to see how this can be done when the basic architectures of two nets differ. The second problem Fodor and Lepore cite is that even if similarity measures for meanings can be successfully crafted, they are inadequate to the task of meeting the desiderata which a theory of meaning must satisfy.

Churchland (1998) shows that the first of these two objections can be met. Citing the work of Laakso and Cottrell (2000) he explains how similarity measures between activation patterns in nets with radically different structures can be defined. Not only that, Laakso and Cottrell show that nets of different structures trained on the same task develop activation patterns which are strongly similar according to the measures they recommend. This offers hope that empirically well defined measures of similarity of concepts and thoughts across different individuals might be forged.

On the other hand, the development of a traditional theory of meaning based on similarity faces severe obstacles (Fodor and Lepore 1999), for such a theory would be required to assign

sentences truth conditions based on an analysis of the meaning of their parts, and it is not clear that similarity alone is up to such tasks as fixing denotation in the way a standard theory demands. However, most connectionists who promote similarity based accounts of meaning reject many of the presupposition of standard theories. They hope to craft a working alternative which either rejects or modifies those presuppositions while still being faithful to the data on human linguistic abilities.

Calvo Garzon (2003) complains that there are reasons to think that connectionists must fail. Churchland's response has no answer to the collateral information challenge. That problem is that the measured similarities between activation patterns for a concept (say: grandmother) in two human brains are guaranteed to be very low because two people's (collateral) information on their grandmothers (name, appearance, age, character) is going to be very different. If concepts are defined by everything we know, then the measures for activation patterns of our concepts are bound to be far apart. This is a truly deep problem in any theory that hopes to define meaning by functional relationships between brain states. Philosophers of many stripes must struggle with this problem. Given the lack of a successfully worked out theory of concepts in either traditional or connectionist paradigms, it is only fair to leave the question for future research.

9. Connectionism and the Elimination of Folk Psychology

Another important application of connectionist research to philosophical debate about the mind concerns the status of folk psychology. Folk psychology is the conceptual structure that we spontaneously apply to understanding and predicting human behavior. For example, knowing that John desires a beer and that he believes that there is one in the refrigerator allows us to explain why John just went into the kitchen. Such knowledge depends crucially on our ability to conceive of others as having desires and goals, plans for satisfying them, and beliefs to guide those plans. The idea that people have beliefs, plans and desires is a commonplace of ordinary life; but does it provide a faithful description of what is actually to be found in the brain?

Its defenders will argue that folk psychology is too good to be false (Fodor 1988, Chapter 1). What more can we ask for the truth of a theory than that it provides an indispensable framework for successful negotiations with others? On the other hand, eliminativists will respond that the useful and widespread use of a conceptual scheme does not argue for its truth (Churchland 1989, Ch. 1). Ancient astronomers found the notion of celestial spheres useful (even essential) to the conduct of their discipline, but now we know that there are no celestial spheres. From the eliminativists point of view, an allegiance to folk psychology, like allegiance to folk (Aristotelian) physics, stands in the way of scientific progress. A viable psychology may require as radical a revolution in its conceptual foundations as is found in quantum mechanics.

Eliminativists are interested in connectionism because it promises to provide a conceptual foundation that might replace folk psychology. For example Ramsey *et al.* (1991) have argued that certain feed-forward nets show that simple cognitive tasks can be performed without employing features that could correspond to beliefs, desires and plans. Presuming that such nets are faithful to how the brain works, concepts of folk psychology fare no better than do celestial spheres. Whether connectionist models undermine folk psychology in this way is still controversial. There are two main lines of response to the claim that connectionist models support eliminativist conclusions. One objection is that the models used by Ramsey *et al.* are feed forward nets, which are too weak to explain some of the most basic features of cognition such as short term memory. Ramsey *et al.* have not shown that beliefs and desires must be absent in a class of nets adequate for human cognition. A second line of rebuttal challenges the claim that features corresponding to beliefs and desires are necessarily absent even in the feed forward nets at issue (Von Eckardt 2005).

The question is complicated further by disagreements about the nature of folk psychology. Many philosophers treat the beliefs and desires postulated by folk psychology as brain states with symbolic contents. For example, the belief that there is a beer in the refrigerator is thought to be a brain state that contains symbols corresponding to beer and a refrigerator. From this point of view, the fate of folk psychology is strongly tied to the symbolic processing hypothesis. So if connectionists can establish that brain processing is essentially non-symbolic, eliminativist

conclusions will follow. On the other hand, some philosophers do not think folk psychology is essentially symbolic, and some would even challenge the idea that folk psychology is to be treated as a theory in the first place. Under this conception, it is much more difficult to forge links between results in connectionist research and the rejection of folk psychology.

10. Predictive Coding Models of Cognition

Predictive coding is a well-established information processing tool with a wide range of applications. It is useful, for example, in compressing the size of data sets. Suppose you wish to transmit a picture of a landscape with a blue sky. Since most of the pixels in the top half of your image are roughly the same shade, it is very inefficient to record the color value (say Red: 46 Green: 78 Blue: FF in hexadecimal) over and over again for each pixel in the top half of the image. Since the value of one pixel strongly predicts the value of its neighbor, the efficient thing to do is record at each pixel location, the difference between the predicted value (an average of its neighbors) and the actual value for that pixel. (In the case of representing an even shaded sky, we would only need to record the blue value once, followed by lots of zeros.) This way, major coding resources are only needed to keep track of points in the image (such as edges) where there are large changes, that is points of "surprise" or "unexpected" variation.

It is well known that early visual processing in the brain involves taking differences between nearby values, (for example, to identify visual boundaries). It is only natural then to explore how the brain might take advantage of predictive coding in perception, inference, or even action. (See (Clark, 2013) for an excellent summary and entry point to the literature.) There is wide variety in the models presented in the predictive coding paradigm, and they tend to be specified at a higher level of generality than are connectionist models so far discussed. Assume we have a neural net with input, hidden and output levels that has been trained on a task (say face recognition) and so presumably has information about faces stored in the weights connecting the hidden level nodes. Three features would classify this net as a predictive coding (PC) model. First, the model will have downward connections from the higher levels that are able to predict the next input for that task. (The prediction might be a representation of a generic face.) Second, the data sent

to the higher levels for a given input is not the value recorded at the input nodes, but the difference between the predicted values and the values actually present. (So in the example, the data provided tracks the differences between the face to be recognized and the generic face.) In this way the data being received by the net is already preprocessed for coding efficiency. Third, the model is trained by adjusting the weights in such a way that the error is minimized at the inputs. In other words, the trained net reduces as much as possible the "surprise" registered in the difference between the raw input and its prediction. In so doing it comes to be able to predict the face of the individual to be recognized to eliminate the error. Some advocates of predictive coding models suggest that this scheme provides a unified account of all cognitive phenomena, including perception, reasoning, planning and motor control. By minimizing prediction error in interacting with the environment, the net is forced to develop the conceptual resources to model the causal structure of the external world, and so navigate that world more effectively.

The predictive coding (PC) paradigm has attracted a lot of attention. There is ample evidence that PC models capture essential details of visual function in the mammalian brain (Rao and Ballard, 1999; Huang and Rao, 2011). For example, when trained on typical visual input, PC models spontaneously develop functional areas for edge, orientation and motion detection known to exist in visual cortex. This work also raises the interesting point that the visual architecture may develop in response to the statistics of the scenes being encountered, so that organisms in different environments have visual systems specially tuned to their needs.

It must be admitted that there is still no convincing evidence that the essential features of PC models are directly implemented as anatomical structures in the brain. Although it is conjectured that superficial pyramidal cells may transmit prediction error, and deep pyramidal cells predictions, we do not know that that is how they actually function. On the other hand, PC models do appear more neurally plausible than backpropagation architectures, for there is no need for a separate process of training on an externally provided set of training samples. Instead, predictions replace the role of the training set, so that learning and interacting with the environment are two sides of a unified unsupervised process.

PC models also show promise for explaining higher-level cognitive phenomena. An often-cited example is binocular rivalry. When presented with entirely different images in two eyes, humans report an oscillation between the two images as each in turn comes into "focus". The PC explanation is that the system succeeds in eliminating error by predicting the scene for one eye, but only to increase the error for the other eye. So the system is unstable, "hunting" from one prediction to the other. Predictive coding also has a natural explanation for why we are unaware of our blind spot, for the lack of input in that area amounts to a report of no error, with the result that one perceives "more of the same".

PC accounts of attention have also been championed. For example, Hohwy (2012) notes that realistic PC models, which must tolerate noisy inputs, need to include parameters that track the desired precision to be used in reporting error. So PC models need to make predictions of the error precision relevant for a given situation. Hohwy explores the idea that mechanisms for optimizing precision expectations map onto those that account for attention, and argues that attentional phenomena such as change blindness can be explained within the PC paradigm.

Predictive coding has interesting implications for themes in the philosophy of cognitive science. By integrating the processes of top-down prediction with bottom-up error detection, the PC account of perception views it as intrinsically theory-laden. Deployment of the conceptual categorization of the world embodied in higher levels of the net is essential to the very process of gathering data about the world. This underscores, as well, tight linkages between belief, imaginative abilities, and perception (Grush 2004). The PC paradigm also tends to support situated or embodied conceptions of cognition, for it views action as a dynamic interaction between the organism's effects on the environment, its predictions concerning those effects (its plans), and its continual monitoring of error, which provides feedback to help ensure success.

It is too early to evaluate the importance and scope of PC models in accounting for the various aspects of cognition. Providing a unified theory of brain function in general is, after all, an impossibly high standard. Clark's target article (2013) provides a useful forum for airing complaints against PC models and some possible responses. One objection that is often heard is that an organism with a PC brain can be expected to curl up in a dark

room and die, for this is the best way to minimize error at its sensory inputs. However, that view may take too narrow a view of the sophistication of the predictions available to the organism. If it is to survive at all, its genetic endowment coupled with what it can learn along the way may very well endow it with the expectation that it go out and seek needed resources in the environment. Minimizing error for that prediction of its behavior will get it out of the dark room. However, it remains to be seen whether a theory of biological urges is usefully recast in PC terminology in this way, or whether PC theory is better characterized as only part of the explanation. Another complaint is that the top-down influence on our perception coupled with the constraint that the brain receives error signals rather than raw data would impose an unrealistic divide between a represented world of fantasy and the world as it really is. It is hard to evaluate whether that qualifies as a serious objection. Were PC models actually to provide an account of our phenomenological experience, and characterize the relations between that experience and what we count as real, then skeptical conclusions to be drawn would count as features of the view rather than objections to it. A number of responders to Clark's target article also worry that PC-models count as overly general. In trying to explain everything they explain nothing. Without sufficient constraints on the architecture, it is too easy to pretend to explain cognitive phenomena by merely redescribing them in a story written in the vocabulary of prediction, comparison, error minimization, and optimized precison. The real proof of the pudding will come with the development of more complex and detailed computer models in the PC framework that are biologically plausible, and able to demonstrate the defining features of cognition.

Bibliography

- Aizawa, K., 1994, “Representations without Rules, Connectionism and the Syntactic Argument,”*Synthese*, 101: 465–492.
- ——, 1997, “Explaining Systematicity,” *Mind and Language*, 12: 115–136.
- ——, 1997a, “Exhibiting versus Explaining Systematicity: A Reply to Hadley and Hayward,” *Minds and Machines*, 7: 39–55.
- ——, 2014, “A Tough Time to be Talking Systematicity,” in Calvo and Symons 2014, 77-101.
- Bechtel, W., 1987, “Connectionism and the Philosophy of Mind: an Overview,” *The Southern Journal of Philosophy*, 26 (Supplement): 17–41.
- ——, 1988, “Connectionism and Rules and Representation Systems: Are They Compatible?,”*Philosophical Psychology*, 1: 5–15.
- Bechtel, W., and Abrahamsen, A., 1990, *Connectionism and the Mind: An Introduction to Parallel Processing in Networks*, Cambridge, Mass.: Blackwell.
- Boden, M. and Niklasson, L., 2000, “Semantic Systematicity and Context in Connectionist Networks,” *Connection Science*, 12: 111–142.
- Butler, K., 1991, “Towards a Connectionist Cognitive Architecture,” *Mind and Language*, 6: 252–272.
- Calvo Garzon, F., 2003, “Connectionist Semantics and the Collateral Information Challenge,” *Mind and Language*, 18: 77–94.
- Calvo, P. and Symons, J., 2014, *The Architecture of Cognition: Rethinking Fodor and Pylyshyn’s Systematicity Challenge,* Cambridge: MIT Press 2014.
- Chalmers, D., 1990, “Syntactic Transformations on Distributed Representations,” *Connection Science*, 2: 53–62.
- ——, 1993, “Why Fodor and Pylyshyn Were Wrong: The Simplest Refutation,” *Philosophical Psychology*, 6(3): 305–319.
- Christiansen, M., and Chater, N., 1994, “Generalization and Connectionist Language Learning,”*Mind and Language*, 9: 273–287.
- ——, 1999a, “Toward a Connectionist Model of Recursion in Human Linguistic Performance,”*Cognitive Science*, 23: 157-205.
- ——, 1999b, “Connectionist Natural Language Processing: The State of the Art,” *Cognitive Science*, 23: 417-437.
- Churchland, P.M., 1995, *The Engine of Reason, the Seat of the Soul : a Philosophical Journey into the Brain*, Cambridge, Mass.: MIT Press.

- ——, 1998, “Conceptual Similarity across Sensory and Neural Diversity: The Fodor/Lepore Challenge Answered,” *Journal of Philosophy*, 95: 5–32.
- ——, 1989, *A Neurocomputational Perspective: The Nature of Mind and the Structure of Science*, Cambridge, Mass.: MIT Press.
- Clark, A., 1989, *Microcognition*, Cambridge, Mass.: MIT Press.
- ——, 1993, *Associative Engines*, Cambridge, Mass.: MIT Press.
- ——, 1995, “Connectionist Minds,” in McDonald (1995), 339–356.
- ——, 2013, “Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science,” *Behavioral and Brain Sciences*, 36(3): 1–73, doi: 10.1017?S0140525X12000477.
- Clark, A., and Lutz, R. (eds.), 1992, *Connectionism in Context*, Springer.
- Cotrell G., and Small, S., 1983, “A Connectionist Scheme for Modeling Word Sense Disambiguation,” *Cognition and Brain Theory*, 6: 89–120.
- Cummins, R., 1991, “The Role of Representation in Connectionist Explanations of Cognitive Capacities,” in Ramsey, Stich and Rumelhart (1991), 91–114.
- ——, 1996, “Systematicity,” *Journal of Philosophy*, 93(22): 561–614.
- Cummins, R., and Schwarz, G., 1991, “Connectionism, Computation, and Cognition,” in T. Horgan and J. Tienson (1991), 60–73.
- Davies, M., 1989, “Connectionism, Modularity and Tacit Knowledge,” *British Journal for the Philosophy of Science*, 40: 541–555.
- ——, 1991, “Concepts, Connectionism and the Language of Thought,” in Ramsey *et al*. (1991), 229–257.
- Dinsmore, J. (ed.), 1992, *The Symbolic and Connectionist Paradigms: Closing the Gap*, Hillsdale, NJ: Erlbaum.
- Eliasmith, C., 2007, “How to Build a Brain: From Function to Implementation,” *Synthese*, 159(3): 373-388.
- ——, 2013, *How to Build a Brain: a Neural Architecture for Biological Cognition*, New York: Oxford University Press.
- Elman, J. L., 1991, “Distributed Representations, Simple Recurrent Networks, and Grammatical Structure,” in Touretzky (1991), 91–122.
- Fodor, J., 1988, *Psychosemantics*, Cambridge, Mass.: MIT Press.
- ——, 1997, “Connectionism and the Problem of Systematicity: Why Smolensky's Solution Still Doesn't Work,” *Cognition*, 62: 109–119.
- Fodor, J., and Lepore, E., 1992, *Holism: A Shopper's Guide*, Cambridge: Blackwell.

- ——, 1999, “All at Sea in Semantic Space: Churchland on Meaning Similarity,” *Journal of Philosophy*, 96: 381–403.
- Fodor, J., and McLaughlin, B., 1990, “Connectionism and the Problem of Systematicity: Why Smolensky's Solution Doesn't Work,” *Cognition*, 35: 183–204.
- Fodor, J., and Pylyshyn, Z., 1988, “Connectionism and Cognitive Architecture: a Critical Analysis,”*Cognition*, 28: 3–71.
- Friston, K., 2005, “A Theory of Cortical Responses,” *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 360(1456): 815–836.
- Friston, K. and Stephan, K., 2007, “Free Energy and the Brain,” *Synthese*, 159(3): 417–458.
- Garfield, J., 1997, “Mentalese Not Spoken Here: Computation Cognition and Causation,”*Philosophical Psychology*, 10: 413–435.
- Garson, J., 1991, “What Connectionists Cannot Do: The Threat to Classical AI,” in T. Horgan and J. Tienson (1991), 113–142.
- ——, 1994, “Cognition without Classical Architecture,” *Synthese*, 100: 291–305.
- ——, 1997, “Syntax in a Dynamic Brain,” *Synthese*, 110: 343–355.
- Grush, R., 2004, “The Emulation Theory of Representation: Motor Control, Imagery, and Perception,” *Behavioral and Brain Sciences*, 27: 377–442.
- Guarini, M., 2001, “A Defence of Connectionism Against the Syntactic Argument,” *Synthese*, 128: 287–317.
- Hadley, R., 1994a, “Systematicity in Connectionist Language Learning,” *Mind and Language*, 9: 247–271.
- ——, 1994b, “Systematicity Revisited,” *Mind and Language*, 9: 431–444.
- ——, 1997a, “Explaining Systematicity: A Reply to Kenneth Aizawa,” *Minds and Machines*, 7: 571–579.
- ——, 1997b, “Cognition, Systematicity and Nomic Necessity,” *Mind and Language*, 12: 137–153.
- ——, 2004, “On the Proper Treatment of Semantic Systematicity,” *Minds and Machines*, 14: 145–172.
- Hadley, R., and Hayward, M., 1997, “Strong Semantic Systematicity from Hebbian Connectionist Learning,” *Minds and Machines*, 7: 1–37.
- Hanson, J., and Kegl, J., 1987, “PARSNIP: A Connectionist Network that Learns Natural Language Grammar from Exposure to Natural Language Sentences,” *Ninth Annual Conference of the Cognitive Science Society*, Hillsdale, NJ: Erlbaum, pp. 106–119.

- Hatfield, G., 1991, “Representation in Perception and Cognition: Connectionist Affordances,” in Ramsey *et al*. (1991), 163–195.
- ——, 1991, “Representation and Rule-Instantiation in Connectionist Systems,” in T. Horgan and J. Tienson (1991), 90–112.
- Hawthorne, J., 1989, “On the Compatibility of Connectionist and Classical Models,” *Philosophical Psychology*, 2: 5–15.
- Haybron, D., 2000, “The Causal and Explanatory Role of Information Stored in Connectionist Networks,” *Minds and Machines*, 10: 361–380.
- Hinton, G., 1992, “How Neural Networks Learn from Experience,” *Scientific American*, 267(3): 145–151.
- ——, 1991, “Mapping Part-Whole Hierarchies into Connectionist Networks,” in Hinton (ed.) 1991, 47–76.
- ——, 2010, “Learning to Represent Visual Input,” *Philosophical Transactions of the Royal Society, B*, 365: 177–184.
- Hinton, G. (ed.), 1991, *Connectionist Symbol Processing*, Cambridge, Mass.: MIT Press.
- Hinton, G., McClelland, J., and Rumelhart, D., 1986, “Distributed Representations,” Chapter 3 of Rumelhart, McClelland, *et al*. (1986).
- Hohwy, J., 2012, “Attention and Conscious Perception in the Hypothesis Testing Brain,” *Frontiers in Psychology*, 3(96): 1–14.
- Horgan, T., and Tienson, J., 1989, “Representations without Rules,” *Philosophical Topics*, 17: 147–174.
- ——, 1990, “Soft Laws,” *Midwest Studies in Philosophy*, 15: 256–279.
- ——, 1996, *Connectionism and the Philosophy of Psychology*, Cambridge, Mass.: MIT Press.
- Horgan, T., and Tienson, J. (eds.), 1991, *Connectionism and the Philosophy of Mind*, Dordrecht: Kluwer.
- Hosoya, T., Baccus, S. A., and Meister, M., 2005, “Dynamic Predictive Coding by the Retina,”*Nature*, 436(7): 71–77.
- Huang, Y. and Rao, R., “Predictive Coding,” *Wiley Interdisciplinary Reviews: Cognitive Science*, 2: 580–593.
- Jansen, P. and Watter, S., 2012, “Strong Systematicity Through Sensorimotor Conceptual Grounding: an Unsupervised, Developmental Approach to Connectionist Sentence Processing,” *Connection Science*, 24(1): 25–55.
- Johnson, K., 2004, “On the Systematicity of Language and Thought,” *Journal of Philosophy*, 101: 111–139.

- Jones, M., and Love, B. C., 2011, "Bayesian Fundamentalism or Enlightenment? On the Explanatory Status and Theoretical Contributions of Bayesian Models of Cognition."*Behavioral and Brain Sciences*, 34(4): 169–188.
- Laakso, A., and Cotrell, G., 2000, "Content and Cluster Analysis: Assessing Representational Similarity in Neural Systems," *Philosophical Psychology*, 13: 47–76.
- Macdonald, C. (ed.), 1995, *Connectionism: Debates on Psychological Explanation*, Oxford: Blackwell.
- Matthews, R., 1997, "Can Connectionists Explain Systematicity?" *Mind and Language*, 12: 154–177.
- Marcus, G., 1998, "Rethinking Eliminative Connectionism," *Cognitive Psychology*, 37: 243–282.
- ——, 2001, *The Algebraic Mind*, Cambridge, Mass.: MIT Press.
- McClelland, J., and Elman, J., 1986, "The TRACE Model of Speech Perception," *Cognitive Psychology*, 18: 1–86.
- McClelland, J., Rumelhart, D., *et al*., 1986, *Parallel Distributed Processing*, Volume II, Cambridge, Mass.: MIT Press.
- McLaughlin, B., 1993, "The Connectionism/Classicism Battle to Win Souls," *Philosophical Studies*, 71: 163–190.
- Miikkulainen, R., 1993, *Subsymbolic Natural Language Processing: An Integrated Model of Scripts, Lexicon and Memory*, Cambridge, Mass.: MIT Press.
- Miikkulainen, R. and Dyer, M., 1991, "Natural Language Processing With Modular PDP Networks and Distributed Lexicon," *Cognitive Science*, 15: 343–399.
- Morris, W. C., Cottrell, G. W., and Elman, J., 2000, "A Connectionist Simulation of the Empirical Acquisition of Grammatical Relations," in Wermter and Sun (2000), 175–193.
- Niklasson, L., and van Gelder, T., 1994, "On Being Systematically Connectionist," *Mind and Language*, 9: 288–302.
- Phillips, S., 2002, "Does Classicism Explain Universality?" *Minds and Machines*, 12: 423–434.
- Pinker, S., and Mehler, J. (eds.), 1988, *Connections and Symbols*, Cambridge, Mass.: MIT Press.

- Pinker, S., and Prince, A., 1988, “On Language and Connectionism: Analysis of a Parallel Distributed Processing Model of Language Acquisition,” *Cognition*, 23: 73–193.
- Pollack, J., 1989, “Implications of Recursive Distributed Representations,” in Touretzky (1989), 527–535.
- ——, 1991a, “Induction of Dynamical Recognizers,” in Touretzky (1991), 123–148.
- ——, 1991b, “Recursive Distributed Representation,” in Hinton (1991), 77–106.
- Port, R., 1990, “Representation and Recognition of Temporal Patterns,” *Connection Science*, 2: 151–176.
- Port, R., and van Gelder, T., 1991, “Representing Aspects of Language,” *Proceedings of the Thirteenth Annual Conference of the Cognitive Science Society*, Hillsdale, N.J.: Erlbaum.
- Ramsey, W., 1997, “Do Connectionist Representations Earn their Explanatory Keep?” *Mind and Language*, 12: 34–66.
- Ramsey, W., Stich, S., and Rumelhart, D., 1991, *Philosophy and Connectionist Theory*, Hillsdale, N.J.: Erlbaum.
- Ramsey, W., Stich, S., and Garon, J., 1991, “Connectionism, Eliminativism, and the Future of Folk Psychology,” in Ramsey, Rumelhart and Stich (1991), 199–228.
- Rao, R., and Ballard, D., 1999, “Predictive Coding in the Visual Cortex: A Functional Interpretation of Some Extra-Classical Receptive-Field Effects,” *Nature Neuroscience*, 2(1): 79–87.
- Rhode, D., and Plaut, D., 2004, “Connectionist Models of Language Processing,” *Cognitive Studies*(Japan), 120(1): 10–28.
- Roth, M., 2005, “Program Execution in Connectionist Networks,” *Mind and Language*, 20: 448–467.
- Rumelhart, D., and McClelland, J., 1986, “On Learning the Past Tenses of English Verbs,” in McClelland and Rumelhart *et al*. (1986), 216–271.
- ——, *et al*., 1986, *Parallel Distributed Processing*, vol. I, Cambridge, Mass.: MIT Press.
- Schwarz, G., 1992, “Connectionism, Processing, Memory,” *Connection Science*, 4: 207–225.
- Sejnowski, T., and Rosenberg, C., 1987, “Parallel networks that Learn to Pronounce English Text,”*Complex Systems*, 1: 145–168.

- Servan-Schreiber, D., Cleeremans, A., and McClelland, J., 1991, "Graded State Machines: The Representation of Temporal Contingencies in Simple Recurrent Networks," in Touretzky (1991), 57–89.
- Shastri, L., and Ajjanagadde, V., 1993, "From Simple Associations to Systematic Reasoning: A Connectionist Representation of Rules, Variables, and Dynamic Bindings Using Temporal Synchrony" *Behavioral and Brain Sciences*, 16: 417–494.
- Shea, N., 2007, "Content and Its Vehicles in Connectionist Systems," *Mind and Language*, 22: 246–269.
- Shultz, T. and Bale, A., 2001, "Neural Network Simulation of Infant Familiarization to Artificial Sentences," *Infancy*, 2: 501–536.
- ——, 2006, "Neural Nets Discover a Near-Identity Relation to Distinguish Simple Syntactic Forms,"*Minds and Machines*, 16: 107–139.
- Smolensky, P., 1987, "The Constituent Structure of Connectionist Mental States: A Reply to Fodor and Pylyshyn," *The Southern Journal of Philosophy*, 26 (Supplement): 137–161.
- ——, 1988, "On the Proper Treatment of Connectionism," *Behavioral and Brain Sciences*, 11: 1–74.
- ——, 1991, "Tensor Product Variable Binding and the Representation of Symbolic Structures in Connectionist Systems," in Hinton (1991), 159–216.
- ——, 1995, "Constituent Structure and Explanation in an Integrated Connectionist/Symbolic Cognitive Architecture," in MacDonald (1995).
- St. John, M., and McClelland, J., 1991, "Learning and Applying Contextual Constraints in Sentence Comprehension," in Hinton (1991), 217–257.
- Tomberlin, J. (ed.), 1995, *Philosophical Perspectives 9: AI, Connectionism and Philosophical Psychology*, Atascadero: Ridgeview Press.
- Touretzky, D., 1989, *Advances in Neural Information Processing Systems I*, San Mateo, CA: Kaufmann.
- ——, 1990, *Advances in Neural Information Processing Systems II*, San Mateo, CA: Kaufmann.
- ——, 1991, *Connectionist Approaches to Language Learning*, Dordrecht: Kluwer.
- Touretzky, D., Hinton, G., and Sejnowski, T., 1988, *Proceedings of the 1988 Connectionist Models Summer School*, San Mateo: Kaufmann.
- van Gelder, T., 1990, "Compositionality: A Connectionist Variation on a Classical Theme,"*Cognitive Science*, 14: 355–384.
- ——, 1991, "What is the 'D' in PDP?" in Ramsey *et al*. (1991), 33–59.

- van Gelder, T and Port, R., 1993, “Beyond Symbolic: Prolegomena to a Kama-Sutra of Compositionality,” in V. Honavar and L. Uhr (Eds.), *Symbol Processing and Connectionist Models in AI and Cognition: Steps Towards Integration*, Boston: Academic Press.
- Vilcu, M., and Hadley, R., 2005, “Two Apparent Counterexamples' to Marcus: A Closer Look,”*Minds and Machines*, 15: 359–382.
- Von Eckardt, B., 2003, “The Explanatory Need for Mental Representations in Cognitive Science,”*Mind and Language*, 18: 427–439.
- ——, 2005, “Connectionism and the Propositional Attitudes,” in C. Erneling and D. Johnson (eds.),*The Mind as a Scientific Object: Between Brain and Culture*, New York: Oxford University Press.
- Waltz, D., and Pollack, J., 1985, “Massively Parallel Parsing: A Strongly Interactive Model of Natural Language Interpretation,” *Cognitive Science*, 9: 51–74.
- Wermter, S. and Sun, R. eds., 2000, *Hybrid Neural Symbolic Integration*, Berlin, Springer Verlag.