COGNITIVE SCIENCE
A Multidisciplinary Journal

Cognitive Science (2014) 1–54

ISSN: 0364-0213 print / 1551-6709 online
DOI: 10.1111/cogs.12148

Parallel Distributed Processing at 25: Further Explorations in the Microstructure of Cognition

Timothy T. Rogers,[a] James L. McClelland[b]

[a]*Department of Psychology, University of Wisconsin-Madison*
[b]*Department of Psychology, Stanford University*

Received 29 July 2013; received in revised form 2 April 2014; accepted 9 April 2014

Abstract

This paper introduces a special issue of *Cognitive Science* initiated on the 25th anniversary of the publication of *Parallel Distributed Processing* (PDP), a two-volume work that introduced the use of neural network models as vehicles for understanding cognition. The collection surveys the core commitments of the PDP framework, the key issues the framework has addressed, and the debates the framework has spawned, and presents viewpoints on the current status of these issues. The articles focus on both historical roots and contemporary developments in learning, optimality theory, perception, memory, language, conceptual knowledge, cognitive control, and consciousness. Here we consider the approach more generally, reviewing the original motivations, the resulting framework, and the central tenets of the underlying theory. We then evaluate the impact of PDP both on the field at large and within specific subdomains of cognitive science and consider the current role of PDP models within the broader landscape of contemporary theoretical frameworks in cognitive science. Looking to the future, we consider the implications for cognitive science of the recent success of machine learning systems called "deep networks"—systems that build on key ideas presented in the PDP volumes.

Keywords: Connectionist models; Neural networks; Language; Cognition; Memory; Learning; Cognitive control; Perception

1. Introduction

Cognitive science might succinctly be defined as the effort to answer three questions: (a) What processes support the complex behaviors of intelligent systems? (b) What kinds of representations do such processes operate on? (c) What is the initial basis of such processes and representations, and how do they change with experience?

Correspondence should be sent to Timothy T. Rogers, Department of Psychology, University of Wisconsin at Madison, 524 Brogden Hall, 1202 W Johnson St., Madison, WI 53706. E-mail: ttrogers@wisc.edu

A snapshot of the current state of the field will reveal vigorous discussion of these questions. In many cases, these contemporary controversies have grown from seeds planted over 25 years ago, with the publication of *Parallel Distributed Processing*—the two-volume collection of articles that addressed shortcomings of the then-dominant symbol processing approach, offered a neurally inspired framework for understanding many aspects of cognition, and challenged many commonly held beliefs about the answers to these core questions (Rumelhart, McClelland, and the PDP Research Group, 1986).

This special issue was initiated in the context of the selection of one of us (JLM) to receive the David E. Rumelhart Prize in Cognitive Science in 2010. This provided an opportunity, about 25 years after the publication of the parallel distributed processing (PDP) volumes, to consider the contributions of the framework and to honor Rumelhart, whose breakthrough contributions spearheaded the PDP approach. Rumelhart passed away in 2011, 25 years after the publication of the PDP volumes. Though a few years have now passed since these milestones, the consideration of these contributions is nevertheless timely, since it comes at a moment when exciting breakthroughs in applying deep neural networks to object recognition (Le et al., 2012), speech perception (Graves, Mohamed, & Hinton, 2013), and sentence processing (Socher, Bauer, et al., 2013) have led to renewed interest in the approach.

The papers assembled in this special issue address several central issues in contemporary cognitive science as viewed through the lens of parallel distributed processing. Each paper traces the development of current thinking from earlier work, places the issues in a contemporary context, and provides pointers toward future directions.

Although each article stands on its own, separate consideration of these works would obscure an important characteristic of the PDP approach—specifically that it offers a set of principles, commitments, and practices that apply across many different cognitive domains. In this introductory article, we consider the framework in general and the answers it gives to the core questions raised above. We begin by considering the symbolic paradigm—already known in the mid-1980s as “good old-fashioned Artificial Intelligence”—along with the phenomena that provided the impetus for developing the PDP alternative. We then review the central features of the PDP approach, consider the key areas where PDP has had an impact, and examine the relationship between PDP and other approaches. We end with a brief consideration of the relevance of the framework for the future.

2. Motivations for the parallel distributed processing approach

2.1. The symbolic framework

From the late 1950s through the 1970s, research in cognitive psychology was largely shaped by the view that the mind is a discrete symbol-processing system, similar to a Turing machine or a serial digital computer. The basic elements of cognition were thought to include arbitrary symbols as representational primitives, structured arrangements of such primitives, and rules for manipulating these to allow for new structured arrangements of symbols (Newell, 1990). Cognitive processes were likened to computer programs—ordered lists of rules that, when applied to some symbolic input, would produce the appropriate output. For instance, a process that computes the past tense form of a verb from its present tense form might be described as follows:

```
IF goal is to generate past-tense form of verb X THEN
retrieve phonological segments of X
bind past tense morpheme [ˆd] to end of X
adjust phonology of past tense morpheme based on final segment of X
END
```

In this pseudo-code, there are representational units (like phonological segments of words and the past-tense morpheme) that are stored in the memory, and there are procedures for operating on these (like the binding operation that joins the past-tense morpheme to the end of the word). Running through the steps of this procedure will produce the correct response for the majority of verbs in the language, and it will also produce a plausible response for novel verbs, thereby accounting for productive use of past tense forms (Pinker, 1999). Of course, more elaborate systems of rules are required to capture more complex cognitive processes—for example, procedures that take a sequence of words as input and apply rules to derive a parse tree that describes the grammatical structure of the sentence (Woods, 1970), or procedures that take two-dimensional intensity arrays as input and return the locations and orientations of the surfaces and edges of the objects that gave rise to the intensity array (Marr, 1982).

This symbolic perspective brought with it implications and assumptions that had a profound effect on cognitive theory. Processes were often thought to be *modular*: Each took a particular kind of input and returned a particular kind of output, applying processes that followed their own domain-specific rules and were impervious to other sources of input (Fodor, 1983; Marr, 1982). Processes were also often thought to operate in a strictly sequential fashion, with each successive process waiting for its predecessor to finish and provide its output (Sternberg, 1969). Many models of human cognitive processes from this period were built on the discrete stage assumption. Thus, in one early proposal, visual word recognition was thought to involve one stage of processing that performs a featural analysis of the stimulus; a second that takes the output of the first and produces a string of identified letters; and a third that uses this string of letter identities as the basis for identifying the word (Gough, 1972). Learning was cast as something of a discrete *all-or-nothing* process—a given symbol, rule, or procedure was either stored in the system or it was not, and policies were developed for determining when rules should be added or merged (Anderson, 1975). Also, because the number of possible sets of rules is unbounded, this view brought with it the assumption that a great deal of *knowledge must be innate* to limit the search space. Language acquisition, for example, was assumed to depend on a pre-specified "universal grammar" subject only to parameterization based on the features of a given natural language (Chomsky, 1965). Finally, the neural underpinnings of the mind were thought to be essentially irrelevant to understanding cognitive function. Just as one need not know the electronic workings of computer hardware to understand the steps of a computer program, so too might one understand the mind's software without worrying about its implementational details (Neisser, 1967).

By the late 1970s, however, it was becoming apparent that models built on these assumptions were failing to capture many fairly basic and fundamental aspects of human cognition. These shortcomings, briefly reviewed here, spurred the search for an alternative set of basic principles for cognition. Some of the points are drawn from the pre-1980 literature and were part of the initial motivation for PDP, but more recent findings that underline the motivating issues are also considered.

2.2. *Graded constraint satisfaction and context sensitivity*

One problem with the symbolic paradigm was the observation that cognitive outcomes are constrained by multiple sources of information. The first chapter of the PDP volumes (McClelland, Rumelhart, & Hinton, 1986) provided several examples. Visual object recognition is accomplished by the simultaneous processing of multiple individually inconclusive cues, each of which may be consistent with many different interpretations. When reaching for an object in a cluttered environment, the locations of the various objects, the current posture of the person reaching, the location of the hand, and the shape, weight, and rigidity of the object to be grasped all exert constraints on the trajectory of the reach. Language comprehension and production are simultaneously shaped by syntactic, semantic, lexical, and morphological constraints, and by the setting or context in which language is produced and by discourse and pragmatic constraints. AI researchers (Minsky, 1974; Schank & Abelson, 1977) proposed "frames" and "scripts" to represent the knowledge needed to guide processing (e.g., of a story about an event such as a visit to a restaurant), but Rumelhart and Ortony (1976) argued that the interpretation of everyday episodes and events requires simultaneous reliance on multiple kinds of knowledge—not just one but several scripts or schemata need to be simultaneously considered. In these and other cases, it was difficult to see how the required sensitivity to many different sources of information could be achieved by modular, discrete, and serial symbol processing systems, or with reference to a single frame or script as context.

The assumption that processing is modular—that a given process operates only on its direct inputs and is unaffected by the operations of other modules—made it difficult to capture the contextual effects on processing that had been robustly documented by the late 1970s (Rumelhart, 1977). One example of such an influence is the word-superiority effect: People can identify a letter more accurately when it appears in the context of a familiar word or word-like letter string than when it appears in isolation or in an unwordlike letter string (Reicher, 1969; Wheeler, 1970). If the process that identifies letters must be completed before the string's status as a word is considered, then letter identification should not depend on word context. Rumelhart (1977) argued that such context effects are ubiquitous in perception and cognition and that this ubiquity called for an approach to modeling cognition in which all aspects of processing can simultaneously influence and be influenced by many different sources of information.

2.3. Content sensitivity

Discrete symbolic approaches drew a bright line between representational structure (the arrangement of elements of cognition) and content (the meanings and referents of these elements). The rules thought to manipulate mental structures were thought to be structure rather than content-sensitive (Fodor & Pylyshyn, 1988; Newell & Simon, 1961). This arrangement was meant to account for the remarkable productivity of much of human cognition. In the case of past-tense formation, for instance, the rule creates a structure that binds together the supplied word-stem and the past-tense morpheme (Pinker, 1991), as long as the stem is a member of the syntactic class of verbs, including a potentially infinite set of novel verbs. The productivity of the rule arose because its use did not depend in any important way on the content to which it was applied—so long as word was "tagged" as a verb, the rule could be applied. Likewise, the rules for constructing parse trees and the structural relations that result could be applied to any sentence so long as the grammatical categories of the words were consistent with syntactic constraints, allowing people to process any syntactically well-formed sentence even if its meaning was novel or anomalous (such as Chomsky's famous example "Colorless green ideas sleep furiously"). Some rule sets could be applied recursively ad infinitum, at least in principle, thus allowing an infinite set of possible sentences to be generated or interpreted using a simple and finite set of rules (see Frazier & Fodor, 1978). These arguments were mainly advanced in the context of linguistics, but cognitive scientists tried to draw a more general take-home point: That the generative and flexible nature of many cognitive processes must indicate the existence of underlying "rules" and abstract relational structures that existed independent of the particular content to which they were applied (Marcus, 2001).

A central problem for this approach, however, arose immediately when computational linguists tried to use it to actually parse sentences. It became apparent that, in many cases, the role assignments of nouns are strongly influenced by word meanings and by the plausibility of the resulting event scenarios that the objects to which the sentence referred might take on. For instance, in sentences like:

The birdwatcher saw the bird with binoculars.

human judgments indicate that the phrase "with binoculars" is interpreted syntactically as attaching to the verb phrase of the sentence and semantically as the instrument of the action of seeing, but in the sentence:

The bird saw the birdwatcher with binoculars.

the phrase "with binoculars" is usually interpreted syntactically as a constituent of the object noun phrase ("the birdwatcher with binoculars") and semantically as an object in the birdwatcher's possession (Oden, 1978). The outcome depends on semantic knowledge about objects and their possible relations in events—in this case, the knowledge that birdwatchers and not birds use binoculars. Oden (1978) and Taraban and McClelland (1988) demonstrated that changes to the verb, the agent, and even distance from the subject to the object could affect such attachment decisions, and similar findings have been reported for many other syntactic decisions. Thus, the structures that result do not appear to be independent of the meanings of the words or of the events the sentences describe. What was needed in language and in other domains was a way of understanding how behavior could be simultaneously sensitive to familiar content while still sufficiently sensitive to structure to allow successful processing of novel, and even semantically anomalous, inputs.

2.4. Sensitivity to quasiregular structure

The idea that cognitive flexibility and generativity arises from the application of rules and abstract structures faced additional challenges. Perhaps most obviously, no matter how rule-like behavior may seem in some domain, there are always exceptions. A canonical example again comes from past-tense formation in English (Pinker, 1999). The standard rule produces the correct past-tense form for most verbs, but there are many others for which it fails (e.g., the past tense form of “go” is not “goed” but “went”). To accommodate such inconsistencies, discrete symbolic approaches typically proposed that, in addition to the rules and structures that govern most behavior, there exists a memory-based system for storing information about exceptional or irregular items. Any given item is then processed either by the rule or the memory system.

In general, however, a sharp distinction cannot be drawn between regular and irregular items. For one thing, many exceptional items come in clusters that can be characterized by patterns of phonological (Bybee & Slobin, 1982) and semantic (Kopcke, 1988) similarity. Attempts to show that languages make a categorical distinction between “algebra-like” rule-following cases on the one hand and similarity-based patterns on the other (e.g., Marcus, Brinkmann, Clahsen, Wiese, & Pinker, 1995) have not been successful (Albright & Hayes, 2003; Hahn & Nakisa, 2000). Even more important is the fact that items can vary in the extent of their participation in and contribution to the regular pattern in a given domain. In past-tense inflection, for example, most exceptional forms share features of the regular past tense pattern (McClelland & Patterson, 2002a). Some incorporate a vowel reduction together with the regular past tense inflection (*say-said* and *keep-kept*). Others differ somewhat more but still have considerable overlap with the regular form (e.g., *leave-left, mean-meant, etc.*). Some that are past-tense like in that they end in a short vowel followed by a /d/ or /t/ do not change at all (*hit, cut*) while some with long vowels followed by /d/ or /t/ simply reduce the vowel to form the past tense (*hide-hid).* Even past-tense forms unrelated to the base verb (*go-went)* contain some features of the “add /d/ or /t/” past-tense schema (as do many other cases, such as *bend/bent, send/sent*). This kind of structure—in which so-called exceptional items participate in the regular patterns of a language *to some degree*—has been described as *quasiregular* structure (Plaut, McClelland, Seidenberg, & Patterson, 1996; Seidenberg & McClelland, 1989).

Quasiregularity is not just a curiosity of inflectional morphology—it appears to characterize many different domains of cognition. Derivational morphology also exhibits quasiregular structure (Seidenberg & Plaut, this issue). In reading, the mapping from spelling to sound (and from sound to spelling) is quasiregular: For example, only one letter in the exception word PINT deviates from typical spelling sound correspondences. Indeed, partial conformity to regular spelling-sound correspondence rules arguably applies to *all* so-called exceptions to English "rules" of orthography. Similarly, many natural kinds share typical properties with many others but have one or two atypical properties, such as the floppy ears of an elephant or the camel's hump (Rogers, Lambon Ralph, Hodges, & Patterson, 2003). In language as in nature, quasiregularity appears to be the rule, not the exception!

In quasiregular domains the division of items into regular versus irregular forms seems both arbitrary and inefficient—arbitrary because there is no principled way to decide how "irregular" an item must be before it is handed off to the functionally independent "exception" system; inefficient because, once handed off, those aspects of an item's structure that are regular must be processed outside the regular system. Instead what one wants is a system that can exploit varying degrees of regularity while still tolerating deviations from the overall structure—a framework in which all linguistic forms are influenced by knowledge of regularities, while also being open to item-specific and similarity-based influences.

2.5. *Learnability and gradual developmental chang*e

Cognitive abilities change dramatically through the first decades of life. What do these changes depend on and how do they occur? Central to the classical framework of cognitive theory is the idea that each level of cognitive ability reflects knowledge in the form of rules, and that change depends on the acquisition or "discovery" of new rules. The discovery of a new rule is often thought to involve a sudden change or "aha moment" (Pinker, 1999), before which the child lacks the rule and after which it is suddenly available.

Because many different systems of rules may be compatible with any set of observations, especially if these observations are restricted to positive examples (instances of members of a category or of grammatical sentences, for instance), it has been standard in classical theory to posit an initial set of proto-rules for each cognitive domain in which young children demonstrate competency (Carey, 2000; Chomsky, 1965; Keil, 1989; Spelke, Breinlinger, Macomber, & Jacobson, 1992). And because rules are discrete, there was a tendency to posit that change involves relatively discrete, sudden transitions from one pattern of performance to another, when a new rule is discovered. This connects naturally with stage theories of development, in which cognition is thought to transition through a series of distinct "stages" which might be pervasive across content domains as proposed by Piaget (1970, 2008) or domain specific (Siegler, 1981). For symbolic theories, the shift from one developmental stage to another reflects the sudden acquisition of a new ability, captured in the discovery of a new rule.

There are problems with this view, however, in both its claims about innate domain-specific constraints and its portrayal of developmental trajectories. With regard to innateness, one problem is the existence of many cognitive domains for which evolution may not have prepared us. Alphabetic writing systems have long been held up as an example, and the discovery of cultures lacking counting systems has suggested that the ability to count is another key example (Gordon, 2004). In both domains, just as in other aspects of language and cognition, the relevant behaviors (reading, counting) are well described by rules to at least a rough approximation. If evolution did not provide the basis for these abilities then they provide case examples of systematic, rule-like cognitive domains that are not innately specified and so must be learned. If rule-like behavior can be learned in these cases, it is not clear why such behavior cannot be learned in other cases. Thus, while there may be important characteristics of innate human abilities that enable the mastery of alphabetic writing systems and systems of number, specific knowledge in the form of rules may not be built in per se, in these or other domains.

A second problem with innateness arguments is the failure of linguistic universals to hold up robustly across cultures (Evans & Levinson, 2009). Cross-cultural universality is meant to be the hallmark of a shared genetic contribution to language. From a study of 30 languages, Greenberg (1963) proposed 45 mostly syntactical universals hypothesized to be shared by all languages. Detailed studies of many more languages, however, have found exceptions to all but a bare handful of quite general shared properties—for instance, that all languages have nouns and verbs. Far more frequent are what have come to be called linguistic tendencies: patterns that are observed across many languages but differ in a few. The few exceptions prove that the more common patterns are not a necessary consequence of our genetic heritage, and again shift the focus of research toward understanding other forces that might shape language, including learning (see Evans & Levinson, 2009 and responses).

Finally, the sheer number of cognitive domains that children may potentially need to master is so great that the idea of domain-specific preparation seems unwieldy. What appears to be needed is the ability to learn, with or without prior constraints that may help to guide initial learning.

With regard to developmental trajectories, the central problem with earlier rule-based accounts is the empirical observation that cognitive change in childhood, though marked by periods of relative stability punctuated by periods of change, is not as abrupt as protagonists of discrete symbol-processing theories (e.g., Marcus et al., 1992) have claimed. In fact, Brown (1973), drawing on Cazden (1968) in his book *A First Language,* noted that children gradually acquire the use of the various inflectional morphemes of English, increasing the probability of using such forms and the range of contexts in which they use them over a period of about a year (McClelland & Patterson, 2002b). Similarly, consider children's performance in Piaget's balance-scale task. Children are shown a scale on which 1–5 weights are placed on a peg located 1–5 steps to either side of a fulcrum and are asked to predict which side of the scale should go down. Children make remarkably consistent errors at different ages, and their behavior has been described by an increasingly elaborate set of rules (Siegler, 1976). Careful studies of developing children, however, subsequently found that children did not transition abruptly from one stage to another, as though they had suddenly discovered a new rule. Instead, periods of stable behavior were punctuated by unstable, probabilistic, and graded patterns of change

(Siegler & Chen, 1998). Children who appear to employ a more complex rule one day might revert the next day to a simpler rule, and a given child might make choices consistent with different rules in the same session. Across these transition periods, behavior improves gradually over time and can exhibit graded sensitivity to previously neglected cues or a wider range of contexts that gradually increases over several years (Shirai & Andersen, 1995; Wilkening & Anderson, 1980). This picture of gradual, graded, and probabilistic change between periods of stable behavior appears to be inconsistent with the discrete acquisition of new rules, as Brown (1973) pointed out. Recent evidence suggests a similar picture for children's mastery of counting principles (Davidson, Eng, & Barner, 2012). In sum, the range of domains in which lawful patterns of behavior are observed and the patterns of developmental change that are generally observed call for a mechanistic theory of learning that does not depend crucially on initial domain knowledge or on the sudden acquisition of rules of a discrete or categorical nature.

2.6. Graceful degradation

The investigation of effects of brain damage on cognition provided another source of motivation for the PDP approach. Patterns of dissociation observed across different cases are often striking, and they have led some to seek pure cases that were thought to reveal most clearly the modular structure of cognitive machinery in the brain (Caramazza, 1984). However, patients with putative disruption of some hypothesized cognitive module —the module for storing knowledge about animals, or for forming the regular past tense, or for reading regular words aloud, to list just three examples—rarely perform completely at floor in the impaired domain, and they rarely perform at ceiling in nominally spared domains (Shallice, 1988). Thus, damage does not generally give rise to either as complete or as specific a loss of function as one might expect in a truly modular system. Instead, brain damage appears to cause *graceful degradation*: graded, probabilistic deficits, with some sparing of function, and with performance strongly influenced by the frequency or familiarity of the stimulus and/or its degree of consistency with other items. Under the modular perspective, lack of specificity was easy to explain: Brain damage is not constrained to affect single modules but might simultaneously compromise several. The graceful degradation of behavior within a single putative module was, however, harder to reconcile with rule-based modular processes. It was not clear how or why damage to a given module might allow the module to continue to perform its task adequately for some items but not for others. What was needed was a means of understanding how damage to a processing mechanism could still allow for partial sparing of performance, even if there was a degree of specificity of the effect.

2.7. Inspiration from neuroscience

Researchers in the symbolic tradition were, of course, aware of many of these phenomena and sought ways of making rule-based systems more graded, probabilistic, and sensitive to multiple sources of information (Anderson, 1983; Laird, Newell, & Rosenbloom, 1987). A final motivation that differentiated the PDP approach from these efforts stemmed from the suspicion that brains might compute quite differently from conventional computers. Perhaps, by paying attention to the way brains actually compute, it might be possible to find computational solutions to address the challenges facing discrete symbol processing approaches we have outlined above. While Neisser (1967) and others steeped in the symbol-processing framework had argued that the details of the hardware implementing cognitive functions was irrelevant, Rumelhart (1989) offered a counter to that view:

> Implicit in the adoption of the computer metaphor is [the assumption that] that we should seek explanation at the program or functional level rather than the implementational level. It is thus often pointed out that we can learn very little about what kind of program a particular computer may be running by looking at the electronics. In fact we do not care much about the details of the computer at all; all we care about is the particular program it is running. If we know the program, we know how the system will behave [...]. This is a very misleading analogy. It is true for computers because they are all essentially the same. [...]. When we look at an essentially different architecture, we see that the architecture makes a good deal of difference. It is the architecture that determines which kinds of algorithms are most easily carried out on the machine in question. It is the architecture of the machine that determines the essential nature of the program itself. It is thus reasonable that we should begin by asking what we know about the architecture of the brain and how it might shape the algorithms underlying biological intelligence and human mental life. (p. 134)

Thus, a key factor in the development of PDP was the exploration of the possibility that the characteristics of human cognitive abilities listed above, all of which pose challenges to the view that the mind is a discrete symbol processing machine like a digital serial computer, might be addressed by exploring the computational capabilities of systems inspired not by the digital computer but by the characteristics of neural systems. Many properties of neural systems seem well suited to addressing the challenges. Neurons can be viewed as comparatively simple processors that integrate information from many different sources, providing a potential mechanism for understanding behaviors that appear to be shaped by many different constraints simultaneously. They operate in parallel and are often reciprocally connected, so that large sets of neurons can all mutually constrain one another as required by any approach to graded constraint satisfaction, context effects, and sensitivity to both structure and content. Neural responses are intrinsically noisy and dependent on the simultaneous participation of large populations of individual neurons, and synaptic connections between neurons can vary in their strength, hinting at an account of the inherently graded and variable nature of both healthy and disordered behaviors. The synapses that connect neurons determine how activity can flow through a neural system, so that changes to these connections can provide a candidate mechanism for understanding learning and developmental change. Together these observations spurred the search for an alternative theoretical approach to cognition that, like symbolic approaches, was both computationally rigorous and explicit in its mechanistic commitments, but that drew its inspiration from neuroscience rather than classical artificial intelligence.

3. The resulting framework

The theoretical framework that arose from these efforts—parallel distributed processing or PDP—differs radically from symbolic approaches to cognition. Perhaps the most salient distinction is that PDP views cognition as emergent (Rumelhart, Smolensky, McClelland, & Hinton, 1986). According to the principle of emergence, the regular or law-like behavior of a complex system is the consequence of interactions among constituent elements, each behaving according to comparatively simple principles that may bear little or no obvious relation to the behavior of the system as a whole (Johnson, 2001; Lewes, 1879). An elegant example is the regular hexagonal structure of the honeycomb, which represents an optimal partitioning of a volume into a set of cells that wastes no space while providing each insect with the most room possible (Elman et al., 1996). Given the elegance and regularity of the honeycomb's "design," a theorist might be tempted to suppose that each bee possesses a set of rules for constructing the honeycomb. In fact, however, the insects simply nestle together and extrude a viscous substance—beeswax—that surrounds its body. The pressure of extrusions from surrounding bees molds the beeswax into its hexagonal shape as an emergent consequence. The elegance and regularity of the resulting structure need not be known to the bee.

According to the PDP approach, cognitive processes and representations are like beehives: They often have quite regular and elegant structure that may appear to reflect accordance with some set of rules or principles of design, yet these properties may emerge from the interactions of simple elements, each behaving according to principles that bear little resemblance to those apparent in the overall behavior of the system. Symbolic or rule-based approaches to cognition may sometimes provide useful approximate descriptions of the system-level behavior, but they may not capture the underlying processes and thereby may fail to capture important characteristics of the system's behavior. The PDP framework provides tools for understanding, not only how apparently systematic and lawful behavior can arise from simpler elements but also the characteristics enumerated above that symbolic cognitive systems lack (McClelland et al., 2010).

3.1. A bird's-eye view of PDP models

What are the "simpler elements" that give rise to emergent behavior, and what are the principles by which they operate? Fundamentally, the elements are assumed to be neurons and synapses. PDP models are not formulated at the single neuron level, however, but capture what are thought to be key aspects of neural processing in simplified form, abstracting away from many of the details of real neural systems. PDP models are thus constructed from an ensemble of *neuron-like processing units* that are best construed as providing a basis for approximating, with a small to moderate number of units (tens to thousands), the kinds of informational states that may exist in much larger populations (up to tens of billions) of real neurons (Smolensky, 1986). Each unit adopts an *activation state* that is a function of its inputs. Units transmit information about their current activation to other units via weighted synapse-like *connections*. Often the units in a model are organized into *layers* that determine the gross pattern of connectivity existing in the network. *Visible units* interface with the environment, allowing inputs to the network or outputs specifying observable responses. *Hidden units* are units whose activations are not influenced directly by external events and do not directly produce observable responses—they only receive inputs from and project outputs to other units. Two examples of PDP models are shown in Fig. 1. In the following text, we describe the general characteristics of such models; the figure and its caption provide additional details, including some of the key equations that govern processing and learning in PDP models.

Processing begins with the presentation of an input, which sets the activations of some of the input units in the network. Each unit then carries out a simple computation: It adjusts its activation in response to inputs from other units via weighted connections. Positive connection weights produce excitatory influences, and negative weights produce inhibitory influences; the magnitude of the weight determines the strength of the influence. Typically these influences are combined to produce a *net input* value which may include a noise term to capture variability in neural activity. An activation function then determines the activation of the unit, which may be binary (on or off with some probability) or real-valued, based on its net input. Although simulations require activations to be updated in a series of discrete steps, this is generally intended to approximate a process in which all units are continuously updating their activations as the activations of other units change.

In some models, connections and their weights are specified directly by the modeler, and hidden units stand for entities designated by the modeler, such as words, concepts, or specified features of items. In learning models, the values of connection weights are determined by a learning rule. In such cases, input and output units often correspond to items or features specified by the modeler, but hidden units in the network have no such pre-specified designation. The functioning of the system and the representations that arise over the hidden units then depend on the learning rule, as well as the connectivity, activation function, and initial conditions of the network. Interest may focus not only on the behavior of the network after a period of learning but also on the time course of learning and on changes in the network's behavior as it learns.

Learning rules in PDP models can generally be characterized as adjusting connection weights to optimize the network's performance as specified by an objective function. Often (as in the function shown in Fig. 1), this function measures the extent to which the activations produced by the network deviate from target values specified in the training environment. The objective function can also, either additionally or alternatively, incorporate sensitivity to correlated activity in the network inputs and can include terms reflecting the magnitudes of the weights or unit activations. In any case, the actual connection adjustment rule always reflects the assumption that learning optimizes the objective function, so it is the objective function rather than the learning rule per se that is critical.

One well-known neural network learning algorithm (called *back-propagation,* Rumelhart, Hinton, & Williams, 1986) has been widely used in PDP models. This algorithm changes each connection weight in the direction that reduces the discrepancy between the network's output and a teaching signal, as represented in an objective function like that in Fig. 1. The computation that determines the correct direction of change for each weight can be viewed as involving the propagation of error signals either backward through connection weights (in a feedforward network, Fig. 1, left) or backward through time (in a recurrent network, Fig. 1, right), hence the algorithm's name. While the method is powerful, it also has limitations as discussed below. Back propagation is, however, only one of many learning algorithms; some others are described by Hinton (this issue).

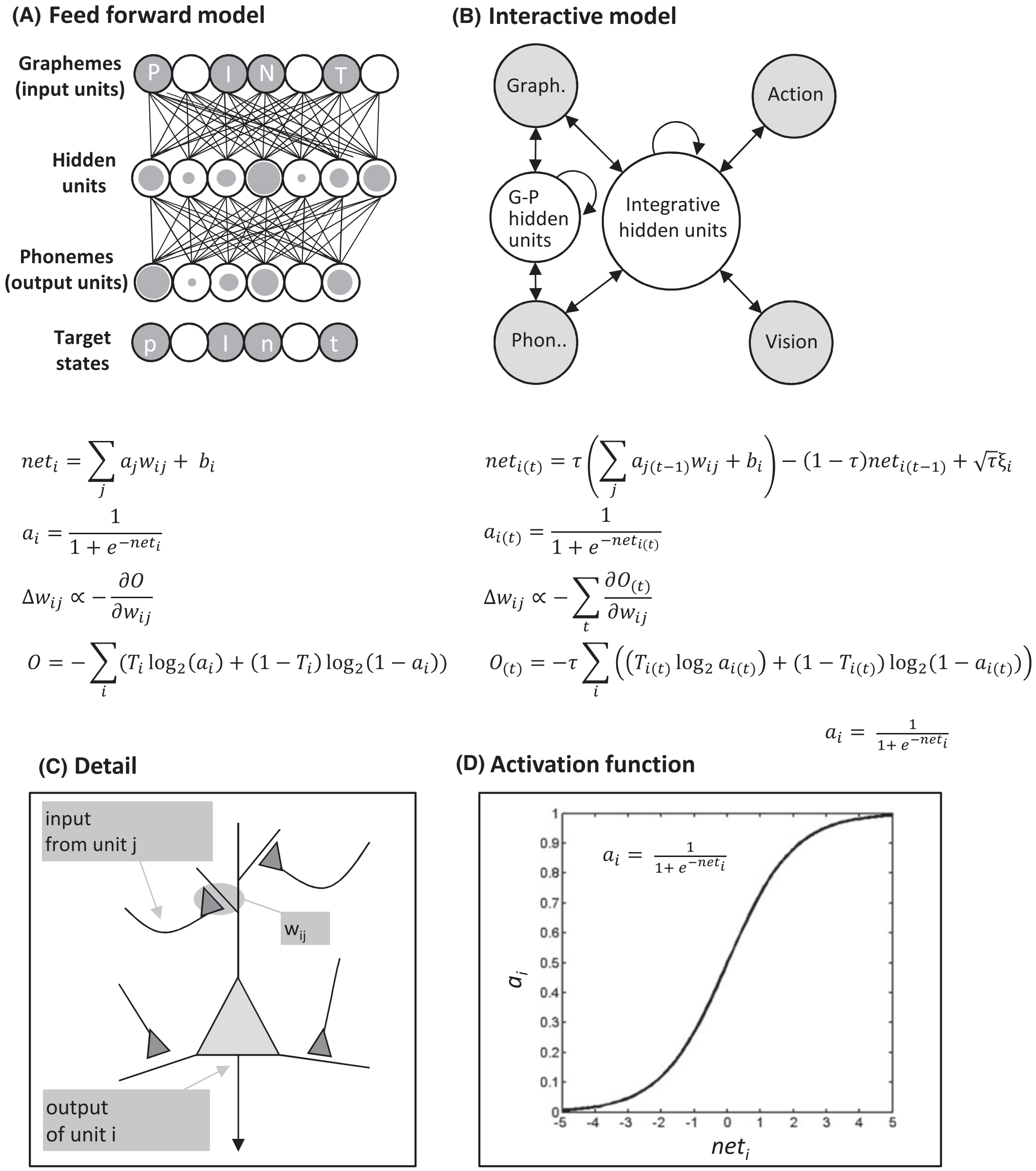
A further crucial element of a learning model is the environment from which it learns. This generally consists of an ensemble of items, each specifying inputs to the network and sometimes target values for outputs from the network or reward signals that depend on network outputs. In some cases the network's outputs can influence the environment, thereby affecting subsequent inputs. In any case, the network learns from exposure to items from its environment, adjusting connection weights based on activation and possibly error signals or reward signals.

3.2. Central tenets of the framework

We now consider the core principles or tenets that have guided the development of the PDP framework from its inception. It is important, though, to recognize that not every researcher working within the PDP approach agrees with all of these tenets, and even modelers (such as ourselves) who accept all of these tenets often use simplified models that do not incorporate all of them. Such simplification is necessary, not only for tractability but also for understanding. Any simulation that attempted to capture neural processes in complete detail would not only be impractical, it would also quite likely be impossible to understand, just as a map that fully captured all aspects of a real physical space at full scale would be completely unwieldy and impractical (Rogers & McClelland, 2008a). We discuss the advantages and disadvantages of simplification more fully below.

3.2.1. Cognitive processes arise from the real-time propagation of activation via weighted connections

Under PDP, all cognitive behaviors ultimately arise from the same underlying process: Units adjust their current activation states according to the weighted sum of their inputs from other units via the connections they receive. Different models adopt different policies for aggregating inputs and adjusting unit activations, but in all models the underlying conception is the same: Cognitive processing involves the propagation of activation among units via weighted connections in a neural network.



$$net_i = \sum_j a_j w_{ij} + b_i$$

$$a_i = \frac{1}{1 + e^{-net_i}}$$

$$\Delta w_{ij} \propto -\frac{\partial O}{\partial w_{ij}}$$

$$O = -\sum_i (T_i \log_2(a_i) + (1 - T_i)\log_2(1 - a_i))$$

$$net_{i(t)} = \tau\left(\sum_j a_{j(t-1)} w_{ij} + b_i\right) - (1 - \tau) net_{i(t-1)} + \sqrt{\tau}\xi_i$$

$$a_{i(t)} = \frac{1}{1 + e^{-net_{i(t)}}}$$

$$\Delta w_{ij} \propto -\sum_t \frac{\partial O_{(t)}}{\partial w_{ij}}$$

$$O_{(t)} = -\tau \sum_i \left(\left(T_{i(t)} \log_2 a_{i(t)}\right) + \left(1 - T_{i(t)}\right)\log_2\left(1 - a_{i(t)}\right)\right)$$

$$a_i = \frac{1}{1 + e^{-net_i}}$$

$$a_i = \frac{1}{1 + e^{-net_i}}$$

3.2.2. Active representations are patterns of activation distributed over ensembles of units

In the PDP framework active representations in the mind are thought to correspond to the patterns of activation generated over a set of units. For example, a percept of a visual input is assumed to be represented as a pattern of activation distributed over many neurons in several different brain areas, and each neuron is thought to participate in the representation of many different items. This representational scheme is held to apply to

Fig. 1. Two examples of Parallel Distributed Processing (PDP) networks. In both activation propagates among neuron-like processing units organized into layers as shown. The lines in A illustrate all connections between units while the bi-directional arrows in B indicate that all units in the adjoined layers are reciprocally connected. (A) The feed-forward model (Plaut et al., 1996) learns to map spellings of letter strings onto their sounds, and can be viewed as a simplification of a part of the adjacent interactive model. Inputs to the grapheme units specify the spelling of a letter string. For each hidden unit the net input (net_i) is computed and converted into an activation value a_i according to the equations shown; the process is then repeated for the output units. Note that b_i indicates the bias of unit *i*, a constant added in to the net input. The model learns from a set of training examples, each consisting of an input pattern corresponding to the spelling of a word and a target output pattern corresponding to the word's pronunciation. Learning occurs by adjusting each connection weight w_{ij} (denoting a weight from unit *j* to unit *i*) according to Δw_{ij}, which follows the gradient of the objective function *O* with respect to each connection weight. *O* expresses the deviation between the activation a_i of each output unit and its target value T*i*, specified by the environment. The function given here, called cross-entropy, is one such function; the sum of the squared differences between each unit's activation and target value is an alternative. (B) The recurrent model (Dilkina et al., 2008) incorporates the principles of continuous interactive processing and intrinsic variability that were left out of the feed-forward network for simplicity. This model learns associations among visual properties of objects, associated actions, and spoken and written representations of their names. It captures the hypothetical process by which, for example, the presentation of the letters C, A, and T can give rise to patterns of activation corresponding to the phonemes in the word "cat," and patterns representing aspects of the meaning of the word "cat," including actions associated with cats and the visual properties of cats. The training environment consists of an orthographic, phonological, visual, and action pattern for each of 240 concrete objects. Time is construed as a number of intervals broken up into ticks whose duration corresponds to some fraction τ of an interval and is small enough to approximate a continuous process. Each unit's net input is based in part on its activation from the previous tick as well as on the activations of other units after the previous tick, subject to independent Gaussian noise represented by ξ_i. Processing begins after an input is provided for one of the visible layers depicted in gray (orthography, phonology, vision or action). On each tick, net inputs are calculated for all units and then activations are updated, and a target value may be provided to any of the visible layers. The objective function *O* is calculated on a tick-by-tick basis, and the change made to each weight is the sum over ticks of the gradient of the objective function on each tick with respect to the weight (see McClelland, 2013a,b for details). (C) A graphic detail for a single unit illustrating the architectural elements corresponding to the terms in the equations. (D) One common function by which net inputs are converted into unit activations.

essentially all kinds of cognitive content: Words, letters, phonemes, grammatical structures; visual features, colors, structural descriptions of objects; semantic, conceptual, and schema representations; contents of working memory and contextual information affecting processing of current inputs; speech plans, motor plans, and more abstract action plans—all are thought to take the form of distributed patterns of activation over large neural populations.

3.2.3. Processing is interactive

The PDP framework proposes that processing is interactive: Propagation of activation among units is generally bidirectional. Thus, if a given unit A sends a connection to a second unit B, there will generally be connections from B back to A, either directly or indirectly via other units. As a consequence, PDP models are often dynamic: Processing of a given input does not complete in a single step but evolves over time. As a given unit alters its activation, it changes the inputs it provides to units downstream, which subsequently alter their own activations, with the potential of feeding back to further alter the state of the original unit. Such mutual influences cause units to continually change their activations over time, until the whole network achieves a *steady state*, that is, a state in which all units have adopted the activations specified by their inputs. Such steady states are called *attractors*, because the intrinsic processing dynamics will cause a network to move its current global state toward a nearby attractor state. In interactive networks, representations of inputs correspond to the attractor state into which the system ultimately settles. Attractor dynamics lie at the heart of the ability of PDP models to exhibit mutual constraint satisfaction, and to find global interpretations of inputs that are mutually consistent. The settling process itself provides a candidate mechanism for understanding how behavior unfold over time in the framework.

3.2.4. Knowledge is encoded in the connection weights

Under PDP, the acquired knowledge that allows a system to behave appropriately does not exist as a set of dormant data structures in a separate store but is encoded directly in the network architecture, in the values of the connection weights that allow the system to generate useful internal representations and outputs. The overall network behavior—the internal representations and subsequent outputs generated in response to a given input—depends critically upon the pattern of connectivity (which units send and receive connections to/from which other units) and the values of the connection weights. In this sense, the knowledge is in the connections; the term *connectionist models* (Feldman & Ballard, 1982) refers to this property.

3.2.5. Learning and long-term memory depend on changes to connection weights

A corollary of the observation that knowledge is stored in the network weights is that learning arises from changes to the weight values. Weight changes are governed by the learning rule that describes how each weight should change with the unit activations and/or error signals that arise in the course of processing different inputs. Models can vary in the particular learning rules they adopt, but in all cases, changes in connection weights arise from the patterns of activation that occur in the network as inputs (including target patterns for output units) are processed. Thus, the generation of appropriate behaviors across different stimuli and contexts, or the acquisition of useful internal representations, all depend on the connection adjustment process.

A further corollary is that, in contrast to many other frameworks, copies of mental states are not stored in long-term memory as such. Instead, the long-term memory trace left behind by an experience is the pattern of connection weight changes it produces. On this view, recollection of a prior experience involves the partial, imperfect reconstruction of aspects of the pattern of activation corresponding to the original experience. This recollection will generally be shaped by the pattern of connection weight changes produced by the experience, as well as by the connection changes produced by other experiences, so that long-term memory is always a constructive process dependent on knowledge derived from many different experiences.

We note that the distinctions among processing, representation, and memory, which are clearly demarked under symbolic approaches, become blurred as a consequence of these proposals. Processing involves propagating activation among units (tenet 1). Active representations are the patterns of activity so produced (tenet 2), which depend upon the values of the connection weights (tenet 3). The connection weights are learned as a consequence of the patterns of activation generated in the network (tenet 4). Thus, processing, representation, and memory are all intrinsically related on this view.

3.2.6. Learning, representation, and processing are graded, continuous, and intrinsically variable

Processing, learning, and representation are held to be intrinsically graded and continuous in the PDP framework. Because representations are coded as distributed patterns of activation over units, representations of different items can be arbitrarily similar to one another. It is common to think of each unit in a distributed representation as corresponding to a dimension in a high-dimensional space, so that any given pattern of activation over units corresponds to a point in the space. All potential representations reside somewhere within this space, which provides a natural medium for expressing continuous-valued similarities among different representations. These similarities in turn provide the basic mechanism for generalization in the framework: Items coded with quite similar patterns of activation (residing near one another in the representational space) will tend to evoke similar downstream consequences, and so to produce similar outputs.

Learning is likewise graded and continuous: It arises from changes to connection weight values expressed as continuous real-valued numbers so that change to any given weight can be arbitrarily small. Moreover, the effect of a weight change is to alter the evoked activations across units in a representation—and since the representations themselves are continuous, the effects of learning can likewise be fully continuous.

Processing, learning, and representation are also assumed to be subject to intrinsic variability. As a result, a given input may result in a distribution of alternative outcomes (represented as quite different resulting patterns of activation) corresponding to alternative interpretations of an ambiguous figure, word, or sentence. Even within a particular outcome, activation will generally vary around a central tendency, and the time required to reach an interpretation will also vary. Because learning depends on network activity and because there may be variability in experiences, learned connection weights are also subject to variability. Whereas intrinsic variability might seem problematic, variability in processing, representation, and learning generally increases the likelihood of finding globally consistent interpretations of inputs and leads to improved generalization and more robust performance under damage.

3.2.7. Processing, learning, and representation depend on the statistical structure of the environment

Finally, the statistical structure of the environment is critically important to understanding cognition under the PDP view. Learning arises from the patterns of activation that arise within neural networks, and from the patterns of errors or violated expectations such activations generate in daily experience. In many cases, and in contrast to some other approaches, PDP explanations of cognitive phenomena depend on the nature of the statistical structure present in everyday experience, and how this structure is exploited by learning in PDP systems.

3.3. Capturing emergent structure: An example

With the basic elements and central commitments of PDP in place, we illustrate how an emergentist approach as embodied in the PDP framework can lead to quite different conclusions about basic issues in cognitive science drawing on Jeff Elman's seminal work on emergent representation of structure in language. Elman (1990, 1991) noted that many aspects of linguistic processing require sensitivity to temporal structure at multiple different time scales. For instance, word recognition requires the ability to segment a temporally continuous speech stream; syntactic processes such as subject-verb agreement require the ability to encode and exploit potentially long-distance relations between words; and sentence comprehension requires the ability to note both the ordering and meanings of words in an utterance in order to determine who did what to whom. Under classical psycholinguistic approaches, these abilities require considerable language-specific machinery: lexemes for denoting and storing information about word forms, representations of grammatical classes that carry important information about inflectional morphology and other syntactic information, and parsing trees that describe the overall grammatical structure of the sentence. Elman (1990) demonstrated that many of the same linguistic phenomena could emerge from a very simple processing system that learned to predict upcoming sentence elements based on a learned internal representation of recent preceding elements, without requiring any of these theoretical constructs.

Elman's model, the simple recurrent network or SRN, is illustrated in Fig. 2A. Its inputs represent information currently available in the environment, and its outputs represent the model's prediction about upcoming information as an utterance unfolds over time. Input units send weighted connections to units in a hidden layer, which in turn send connections to the output layer. Finally, the SRN incorporates the idea that the current hidden state can be constrained by its own prior state, as well as the current input. To implement this the model contains an additional layer of units, the *context* layer, which retains a copy of the pattern of hidden layer activations on the previous processing step, and has modifiable connections to the hidden layer, allowing the network to learn how to use temporal context.

Elman (1990) used this framework to illustrate how word boundaries might be discovered in a continuous speech stream, using a network that did not include word forms as representational primitives. Instead, he simply constructed a training corpus based on a simple narrative text, and presented the text to the network one letter at a time. With each letter, the network was trained to predict the subsequent letter. After learning, prediction errors tended to be high at the boundaries between words and low within words, so that prediction error could serve as a signal for segmenting an otherwise continuous stream of letters into

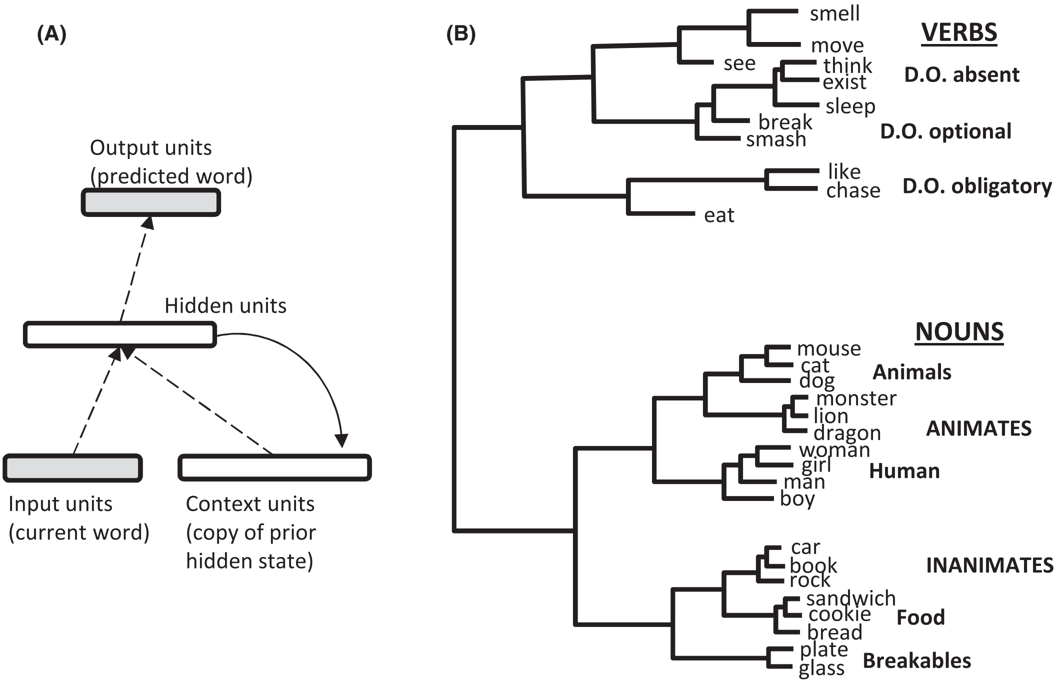


Fig. 2. (A) The simple recurrent network used by Elman (1990). The network is trained with a sequence of patterns corresponding to words. When item *i* in the sequence is presented to the input units, item *i*+1 is used as the target for the output units; this item becomes the input pattern for the next step in the sequence. The activation and learning processes in the network are similar to those of the feedforward network shown in Fig. 1A. The curved solid arrow from the hidden layer to the context layer corresponds to a copy operation, which occurs after each item is processed, so that the hidden unit pattern resulting from processing item i is available as context for processing item *i*+1. Panel B shows a hierarchical clustering analysis of the pattern of activation on the hidden layer for each of the nouns and verbs in the training corpus. The similarities are learned from the temporal structure of the sentences in the corpus and are not due to overt similarities among the input and output patterns for nouns and verbs, which were arbitrarily assigned.

words. Though letters were used for simplicity, the same principles apply for sound sequences. Thus, on this view, word segmentation emerges from sensitivity to the statistical structure of sound sequences in spoken input. Knowledge of the word forms is latent in the connection weights in the network and is not explicitly or transparently represented.

Elman (1990) also used the SRN to illustrate how other linguistic constructs, such as grammatical classes and semantic categories, might emerge from word sequences. Each word was assigned an arbitrary unique pattern, and word sequences were formed by stringing together a series of short sentences obeying simple grammatical and semantic constraints (e.g., action verbs require animate subjects, some action verbs also require direct objects, while others do not; characteristics of objects vary across different action verbs). When the network had learned, the similarity structure of the learned patterns over the hidden units captured information about each word's grammatical class and aspects of its meaning (Fig. 2B). For instance, verbs all tended to elicit grossly similar patterns of activation that were quite different from nouns, and within verbs, transitive verbs tended to cluster together with one another and to differentiate from intransitive verbs. Within the nouns, word representations clustered according to aspects of their meaning: Animate nouns all elicited patterns that were similar to one another and dissimilar to inanimate nouns.

Recall that the network was not explicitly trained to generate semantic or grammatical information, but simply to predict the upcoming word in a sentence. Nevertheless, the internal representations emerging from learning in this task appeared to capture important grammatical and semantic information. What accounts for this phenomenon? The answer is that such information is present in the statistical characteristics of the training environment and is helpful for predicting subsequent words. Training Elman's network to predict each successive word causes it to assign similar connection weights and therefore similar internal representations to items that make similar predictions. Semantic and grammatical structure in the learned representations thus emerges as a consequence of the pressure to optimize prediction. As in the case of word segmentation, however, the structure is not directly or transparently represented—it is latent in the similarity structure of the patterns of activation generated in response to each word, which in turn is the result of a learning process sensitive to the statistical structure of the language. The structure is emergent.

Finally, Elman demonstrated that this simple learning and processing mechanism was capable of explaining an aspect of linguistic processing central to the argument that grammatical processes must be based on a complex system of structure-sensitive rules—specifically, the observation that people can correctly inflect verbs to agree with subjects even in sentences where subject and verb are separated by a sequence of words corresponding to one or more embedded clauses (Elman, 1991). For instance, in this sentence:

The man who owns dogs who chase cats likes ice-cream

the auxiliary verb near the end of the sentence must be the singular "likes" to agree with its syntactic subject "the man," despite the intervening clauses containing plural nouns. Correctly predicting the singular verb seems to require that participants keep track of the hierarchical grammatical structure of the sentence, which flags "the man" as the subject of the main clause and joins this to the verb phrase "likes ice cream." Under the symbolic view, participants apply linguistic rules to construct parsing trees that explicitly assign the relevant constituents to the appropriate roles. But Elman's simulations showed that the relevant constraints can be learned from sentences with embedded clauses exhibiting the appropriate agreement relationships. Thus, the network can acquire knowledge of the grammatical constraints without explicit grammatical rules and without constructing an explicit parse tree; no innate principles of grammar are necessary.

3.4. Simplification and variety in PDP models

With this brief overview of the framework and one example of its use, we are now in a position to revisit questions about the simplifications adopted in PDP models. The models Elman used to investigate aspects of language do not adhere to all of the theoretical commitments stated earlier. For instance, the framework states that processing is continuous, interactive, and intrinsically variable, yet all of Elman's models are essentially deterministic, feed-forward models in which activations are updated only once for each successive input. Moreover, Elman's different simulations vary in how they represent inputs and outputs. In the word-segmentation simulation, the input is a sequence of letters, while in the sentence-level models, words are pre-supposed. In different word-level simulations, input word patterns are sometimes distributed patterns and sometimes involve a unit dedicated to each word. From these examples, and from the highly variable nature of the specific models that all fall under the umbrella of the PDP approach, it is clear that there is significant variation from case to case. What justifies such deviations from the core tenets?

The answer is that a given model is generally constructed to bring out the implications of *a subset* of the tenets of the framework, and, as previously noted, simplification is of the essence in allowing practical progress and understanding. The analogy to cartography introduced above may again be helpful: Maps can be used for many different purposes, and the elements incorporated in a given map depend upon the purpose. For a mountaineer, it is critical to have a map expressing information about changes in elevation. For the traveler who prefers to fly, it is more important that the map indicates the locations of airports and flights that connect them. The best map for a given purpose is the one that strips away distracting details that may nevertheless be critical to other uses. Likewise, any individual PDP model is constructed to serve a particular purpose—to allow the theorist to explore and to illustrate the relationship between a set of hypothesized mechanisms and a particular behavior. Features that are not critical for a particular purpose may be left out, even if these features are part of the general framework and are critical in other contexts.

It must be acknowledged that there are pitfalls, as well as advantages, of such simplification. Which features of a model are essential to its successes? Are its failures intrinsic to the principles at stake or would a less simplified model capture phenomena that a simplified model cannot explain? If different phenomena are addressed by models adopting different simplifications, is an integrated model addressing all of the phenomena simultaneously possible? Will simplified models "scale up" when applied to larger, more complex, and more realistic learning environments? These are legitimate questions that are addressed in different ways in several of the articles in this special issue; the scaling issues are especially important for machine learning applications, which we consider in more detail below (and see Hinton, this issue). In our view, understanding how the phenomena observed in a particular domain might actually emerge from interactions among simple processing units is far from trivial, and each model that has been proposed provides only a partial understanding. We will consider how this understanding may advance further in the future in the final section of this article.

3.5. Relation to rational and computational-level models

It is important to consider how the PDP approach to modeling cognition relates to what Anderson (1990) calls a *rational analysis* and Marr (1982) called a *computational-level* analysis. Anderson and Marr both propose that the attempt to understand a cognitive process should begin with an analysis of the problem to be solved. Once we understand the problem itself, along with the information available to solve it, we can formulate how we should best proceed to discover a solution. Certainly this is a useful approach to artificial intelligence, where the goal is often to find the best possible solution to specific, well-defined problems.

PDP has tended to take a somewhat different approach—framing the computational problems more generally, asking how they might be addressed by brain-like hardware, and only then considering specific problems. As one example, consider the Boltzmann machine (Hinton & Sejnowski, 1986), an early neural network designed to illustrate how a system of neuron-like elements could come to encode, through purely local computations, an overall interpretation of a complex sensory input. The Boltzmann machine was formulated as a solution to a very general problem: That of achieving an optimal interpretation based on partial and/or ambiguous evidence and graded, probabilistic constraints, while relying on the properties of neural hardware. The Boltzmann machine frames an approach to a broad class of specific cases, making it possible to view specific models, such as the multinomial interactive activation model (McClelland, Mirman, Bolger & Khaitan, this issue), as instances of this broader class. More important, the Boltzmann machine learning algorithm (Ackley, Hinton, & Sejnowski, 1985) and its successors (e.g., deep learning methods based on restricted Boltzmann machines, Salakhutdinov & Hinton, 2009) provide a means whereby such optimal models can be learned for particular cases. This contrasts with the approach of Marr (1982), who began by dividing a complex function (such as vision) into several better-defined and highly specific subtasks (such as identifying the surfaces of objects from the pattern of reflected light reaching the retina), then explored solutions to the simpler sub-tasks.

Both approaches have clearly had a strong impact on cognitive science and both have led to important insights. The virtues of optimality based approaches have been frequently described in the literature (Griffiths et al., 2010; Tenenbaum, Griffiths, & Kemp, 2006), but the advantages of the PDP approach are less well known, so we will briefly articulate them here.

First, human perception and action appear to exploit all possible sources of information as they become available in real time. Many optimality analyses consider only what the best outcome should be regardless of the time required to perform the computation. PDP models typically address such constraints, for instance by modeling time dependence of outcomes or by including pressure to respond quickly in the objective function. Indeed, the objective function used in training recurrent neural networks as given by Fig. 1B is often used to implement such a pressure, and as learning progresses responding occurs more and more quickly (Plaut & Booth, 2000). This temporal connection also allows the models to be constrained by a broader range of data, such as the response-time data that are the bread-and-butter of cognitive psychology.

Second, whereas optimality analyses typically consider just the computational problem and the available information, the PDP approach additionally considers how characteristics of the available hardware (neurons and connections) constrain the nature of the solutions found (McClelland et al., 2010). We believe this to be a virtue because adherence to these constraints can lead to new and quite different solutions to the general problems, as noted in the quotation from Rumelhart above. From this perspective, PDP offers a set of general principles by which the mechanisms and structures available to humans operate, within which the constrained optimization carried out by the learning mechanism operates.

Third, a rational analysis requires an explicit representation of the details of the computational problem, but such a characterization may not be straightforward or even appropriate for many problems of great interest. Various explicit characterizations of natural structure have been proposed—hierarchies, directed graphs, clusters of exemplars generated from a set of prototypes, generative grammars, and so on—but we have argued that these only roughly approximate natural environments, which can be quasiregular and subject to a wide range of interacting constraints. In such cases it may be better to allow the structure latent in the environment, together with the constraints built into a network's objective function, to implicitly define the computational problem, and to then allow the learning process to solve the problem by taking the structure as it actually is (see McClelland et al., forthcoming).

Finally, explicit rational or computational-level analyses often pre-suppose a task decomposition that may well not be valid. For example, many models of sentence processing suppose that a syntactic parse tree of an incoming sentence must be formed as a specific stage in the process of understanding, but this proposal has long been contested (e.g. Schank, 1981). PDP models do not completely obviate the problem—they, of course, must also delineate the nature of the task in some way. In many cases, however, models can be formulated by stipulating only task elements that are largely uncontroversial. The task can be posed more directly in terms of deriving appropriate outputs given particular inputs. For instance, in the case of sentence comprehension, a model may take phonological representations of a series of words as input and generate single-word responses to questions about the meaning of the sentence as output (McClelland, St. John, & Taraban, 1989). Since all theories of sentence processing must allow for phonological inputs and judgments about sentence meaning as outputs, these stipulations seem uncontroversial. Intermediate steps assumed by other theories that may be more controversial, such as the construction of an explicit parsing tree, can be omitted. For some tasks, of course, delineating the task solely with respect to uncontroversial elements is very difficult. For instance, Chang, Dell, and Bock (2006) have suggested that PDP models of speech production better capture important aspects of human behavior when role-filling and structure assignment are computed by separate components of the network—a solution that extends the scope of the framework to a challenging domain but relies on a less straightforward and potentially controversial decomposition of the task.

In summary, PDP models implicitly find solutions to posted computational challenges, not by an explicit rational or computational-level analysis but by a constrained optimization process, sensitive to the criteria embodied in the objective function, the affordances of the underlying hardware, and the structure in the environment.

4. Impact of the PDP approach

Having indicated the motivations for the PDP framework and the key elements of the approach, we are now ready to consider its impact. Here we take an avowedly positive stance, laying out some of the ways in which the framework (a) has offered new and productive ways to think about basic issues in cognitive science, (b) has provided fruitful new theories in specific cognitive domains, and (c) has begun to address some of the key computational challenges that motivated the search for alternatives to symbolic approaches. The subsequent section considers critiques, shortcomings, and challenges that remain to be addressed.

4.1. Impact on representation and processing: No rules required

PDP models have been instrumental in illustrating how highly generative, productive aspects of behavior might be supported by mechanisms that do not implement explicit rules. The roots of this work lie in early models of word-reading and past-tense inflection that illustrated how correct responses (the phonological form of a word or the past-tense form of a verb) could be generated from inputs (the visual form of a word or the present-tense form of a verb) for both regular and irregular items within a single system that did not possess separate "rule" and "exception" systems. Reading and inflectional morphology constitute the two domains where arguments supporting rule-based approaches to cognition had been most comprehensively advanced (Pinker, 1999)—so the demonstration that rules were not necessary to explain behavior in these domains served to challenge the broad assumption that much of cognition is rule based. This challenge in turn altered the kinds of questions that language researchers and other cognitive scientists were asking, and the kinds of evidence that entered into the argument. If the interesting aspects of behavior in some domain do not depend on explicit sets of rules, questions about how the rules are learned become moot. So too do accompanying demonstrations that the rules are unlearnable and the resulting conclusion that much rule-based knowledge must be innate. Instead, the nature of learning and memory, the statistical structure of the learning environment, perceptual work in phonology and orthography, and the demands placed on sequencing and articulatory systems, all became relevant to understanding basic aspects of language. The reading and past-tense models thus threatened a kind of domino effect with the potential to topple a fairly wide-ranging set of assumptions about how language specifically, and cognitive science more generally, should be approached. Critical reactions to these ideas spurred the development of more detailed models by both advocates (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001; Perry, Ziegler, & Zorzi, 2010) and critics (Joanisse & Seidenberg, 1999; MacWhinney & Leinbach, 1991; Plaut et al., 1996; Plunkett & Marchman, 1996; Woollams, Lambon Ralph, Plaut, & Nagaraja, 2007) of rule-based approaches to cognition—a trajectory that is reviewed in detail by Seidenberg and Plaut (this issue). Regardless of one's perspective on the resulting theories, however, it is clear that these ideas opened the door to alternative ways of thinking about the mechanisms that support the remarkable flexibility and generativity of human behavior.

4.2. Origins of knowledge and developmental change: Rethinking innateness

A common conclusion from much research in cognitive development has been that mature cognition develops from initial, innately specified and domain-specific knowledge structures. A major contribution of PDP over the past two decades has been to show that, in many cases, evidence thought to support claims of innate domain-specificity is in fact fully consistent with domain-general, experience-based approaches (Elman et al., 1996).

One form of evidence has been the fact that children and adults often behave in a domain-specific manner—drawing quite different conclusions about the properties or behaviors of items from different domains given some observation (Carey & Spelke, 1994; Gelman & Williams, 1998; Wellman & Gelman, 1997). For instance, children treat different properties as "important" for generalizing new information depending upon what "kind of thing" the items are thought to be (Carey, 1985)—extending internal properties (e.g., "has warm blood inside") across items with the same name even if they are different in appearance, but extending physical properties (e.g., "is ten tons") across similar-looking items even if they have different names (Gelman & Markman, 1986). For some researchers, such domain-specific behaviors pose a puzzle for knowledge acquisition: To determine to which domain a given item belongs, it is necessary to know how its observed properties should be weighted, but one cannot know the appropriate feature weightings without knowing the domain (Gelman & Williams, 1998). To resolve the apparent circularity, many theorists have proposed that we are born with innate knowledge of domain-specific constraints that specify which classes, properties, and relationships are important for organizing knowledge in which domains. But the paradox that motivates this reasoning disappears in the PDP framework. PDP models are capable of learning complex statistical relationships among properties of objects and events, and as a consequence can show the emergence of domain-specific patterns of responding from the application of a domain-general learning rule (Colunga & Smith, 2005; Mareschal, French, & Quinn, 2000; Mayor & Plunkett, 2010; Rogers & McClelland, 2004; Westermann & Mareschal, 2012). Thus, the empirical demonstration of domain-specific patterns of responding in children and adults does not provide evidence that there exist innate domain-specific knowledge structures.

The second argument concerns the relatively young ages at which infants can sometimes exhibit traces of domain-specific knowledge. Spelke, for instance, has provided evidence that young infants expect unsupported objects to fall and are surprised when objects pass through a barrier (Spelke et al., 1992). In these cases, however, the knowledge that appears to emerge earliest in development encompasses the most common and systematic structure in the environment. Essentially all physical objects, be they nonliving, inanimate, or animate, behave in accordance with the physical principles Spelke has outlined. Consequently, these are precisely the kinds of things one would expect a domain-general learning system to acquire most rapidly, and so they should be observed earliest in development. Indeed, features that are shared by all of the objects in the domain of living things are learned very rapidly in the Rumelhart model of semantic knowledge acquisition (Rogers & McClelland, 2004).

A third form of evidence is that children often show patterns of behavior over time that others have seen as difficult to reconcile with simple learning-based accounts of developmental change. U-shaped curves in cognitive development provide a dramatic example: The child shows a correct pattern of behavior early in life (e.g., using "went" as the past-tense of "go"), then exhibits incorrect responses that essentially never occur in the environment (e.g., saying "goed" instead of "went"), later eliminating these apparently creative errors (Siegler & Alibali, 2005). It may seem as though experience-driven learning processes could never lead a child to abandon a correct behavior once learned, especially to produce a form that is never actually encountered directly. Such findings have motivated the appeal to some process other than incremental learning, such as the "discovery" of a past-tense rule that "over-rides" the correct response the child had previously learned (Pinker, 1999). PDP models have been integral in demonstrating that this conclusion does not follow. In fact the same learning mechanisms that lead to stage-like patterns of change in some models can also produce U-shaped patterns of change (Plunkett & Marchman, 1991[1]; Rogers, Rakison, & McClelland, 2004). Thus, the U-shaped pattern does not compel the conclusion that there are forces beyond general learning processes at play.

An advantage of the PDP approach to early cognitive abilities and their change over the course of development is the focus this approach provides on explaining, not just how early competencies arise, but how and why cognitive abilities change over time. Nativist approaches to early competencies raise the question of why infants do not always succeed at tasks of interest as soon as they can be tested, and how and why behaviors develop at all. Nativists often suggest that young infants have the cognitive competency to succeed, but lack the performance skills to display this knowledge. However, this proposal is not always wedded to specific or testable claims about what performance characteristics are needed or how these develop. The learning algorithms adopted in PDP theories, in contrast, provide concrete mechanisms that link patterns of developmental change to specific claims about the structure of the environment, and about how experience with that structure will alter processing in the system (Elman et al., 1996; Mareschal, 2000; Munakata & McClelland, 2003; Munakata, McClelland, Johnson, & Siegler, 1997; Plunkett & Sinha, 1992). Because competence is always a matter of degree, partial success in less demanding situations followed by greater robustness naturally emerges within the PDP framework. The PDP approach to these issues overlaps with approaches taken by researchers exploring dynamical systems approaches to development (Spencer, Thomas, & McClelland, 2009; Spivey, 2008; Thelen & Smith, 1996).

4.3. Impact on cognitive neuropsychology: Explanation without the transparency assumption

As noted earlier, the strongly modular cognitive neuropsychology that arose in the 1970s—and in particular, its adoption of the *subtractivity* or *transparency assumption* under which disordered behavior reflects the normal operation of undamaged components absent the contribution of damaged modules—made it difficult to understand the often graded and mixed patterns of impairment that characterize most neuropsychological syndromes. The PDP framework transformed the landscape in several important ways (see Shallice & Cooper, 2011 for detailed review).

First, damaged neural networks exhibit the graceful degradation often observed in brain-damaged individuals: Small amounts of damage have subtle effects, and even with large amounts of damage the system can often generate correct performance with short, frequent, or typical items as well as partially correct responses to other inputs. Some of the factors shown to influence network performance under damage were similar to those that influence the behavior of neuropsychological populations. Indeed, performance with frequent and typical items tends to be spared, as observed across a number of domains in many patient groups (Patterson et al., 2006).

A second important contribution related to the interpretation of double dissociations in neuropsychology. A double dissociation is observed when lesion A affects function 1 more than function 2, while lesion B affects function 2 more than function 1. Such dissociations can, of course, be quite striking, with performance in the spared domain at or near ceiling while performance in the affected domain is very seriously impaired. Classically, such a pattern was treated as implicating separate modules for function 1 and function 2. PDP models, when subjected to simulated damage, can also show very striking patterns of double-dissociation (Plaut & Shallice, 1993), but importantly the dissociated functions need not align with elements of the underlying architecture. For example, Farah and McClelland (1991) illustrated how semantic knowledge about animals and manmade objects can doubly dissociate in a system where items from both domains are represented as distributed patterns of activation over interacting populations of units that encoded perceptual or functional information. Representations of animals had a larger proportion of visual than functional semantic features, whereas the reverse was true for manmade objects. Removal of connections involving the visual semantic units thus produced a greater knowledge deficit for animals than for manmade objects. Importantly, however, the deficit affected both visual and functional attributes of animals: Because the units in the semantic representations interacted with one another, loss of knowledge about visual semantic properties led to catastrophic collapse of the animal representations, so that even the functional information could not be recovered for these items. In contrast, the visual properties of manmade objects often could be retrieved despite this damage, because these properties received input from the intact functional features that made up most of the representation for manmade objects. The reverse pattern was observed when connections to the functional features were damaged. The apparent double dissociation of knowledge for animals versus artifacts did not arise because there were separate parts of the network for processing these different kinds—instead, it arose from a more fundamental specialization for visual and functional information about all kinds of things, coupled with learning about how these kinds of properties tend to covary with one another in the environment. Several other groups have also relied on PDP models to offer explanations for category-specific patterns of deficit that go beyond what could be provided by inferences based on the transparency assumption (Devlin, Gonnerman, Andersen, & Seidenberg, 1998; Lambon Ralph, Lowe, & Rogers, 2007; McRae & Cree, 2002; Tyler, Moss, Durrant-Peatfield, & Levy, 2000).

In the intervening years, these and related ideas have been greatly elaborated. Plaut (2002), building on earlier work by Jacobs & Jordan (1992), explored the possibility that the degree of influence between populations of neurons might partly be a function of their neuroanatomical proximity. Models that incorporate this constraint acquire an emergent, graded form of functional specialization, such that some units contribute more to certain input–output mappings than do others. From this idea, Plaut (2002) developed a computational account of optic aphasia, a puzzling syndrome that has resisted explanation under standard modular assumptions (Farah, 1990). Separately, the Farah and McClelland model has been extended to provide a comprehensive account of both the domain-general semantic impairment observed in semantic dementia (Dilkina, McClelland, & Plaut, 2008; Rogers et al., 2004) and the category-specific impairment often observed in patients with herpes viral encephalitis (Lambon Ralph, Lowe, & Rogers, 2007). Finally, PDP models of single word reading have been able to address the puzzling symptom complex of deep dyslexia (Plaut & Shallice, 1993), as well as effects of semantic deficits on reading (Plaut et al., 1996; Woollams et al., 2007) and past-tense inflection (Joanisse & Seidenberg, 1999). Other influential PDP models have now been developed to explain disordered behavior in domains as diverse as cognitive control (Cohen, Aston-Jones, & Gilzenrat, 2004), working memory (O'Reilly, Braver, & Cohen, 1999), many varieties of aphasia (Ueno, Saito, Rogers, & Lambon Ralph, 2011), and sequential action (Botvinick & Plaut, 2004). In general cognitive neuropsychology has come to rely more on explicit, implemented models of healthy and impaired performance—a trend largely attributable to the successes of PDP models in this area.

PDP approaches have had a similar impact in explorations of developmental disorders associated with abnormal developmental profiles such as Williams syndrome or the so-called Specific Language Impairment. Some have sought explanations for these conditions in terms of the idea that the disorders impact specific modular subsystems, but there are difficulties with this approach (Karmiloff-Smith, 1992), and alternatives based on the possibility that alterations of parameters of neural networks have been articulated within the PDP framework (Thomas & Karmiloff-Smith, 2003).

4.4. Impact on machine learning

Neural networks that preceded and later grew out of ideas described in the PDP volume have always played an important role in the field of machine learning, an outgrowth of the field of artificial intelligence. While much of the motivation for the PDP approach came from psychological considerations, the idea that artificial neural networks that learn graded constraints from experience might provide the best way to emulate human capabilities in a wide range of domains has deeper roots, originating with work by McCullough and Pitts in the 1940s and reaching an early heyday in the 1950s with Rosenblatt's perceptron (Rosenblatt, 1958) and Samuel's (1959) program that learned to play the game of checkers. Disenchantment with the perceptron set in during the 1960s (Minsky & Papert, 1969), but the emergence of back-propagation and related neural network learning methods in the early 1980s revived interest and fed a wave of excitement reflected in the creation of new communities such as the Neural Information Processing Systems community in 1987. Much of the excitement depended on the fact that back-propagation afforded the opportunity to train all layers of a multi-layer network. Excitement about the prospects of multi-layer neural networks waned a second time during the 1990s (see Hinton, this issue, for discussion) but several machine learning groups kept working with neural network approaches throughout, and exciting new directions in what has come to be called deep learning (i.e., learning in multi-layered neural networks) emerged in the early years of the current century (Bengio, Lamblin, Popovici, & Larochelle, 2007; Hinton & Salakhutdinov, 2006; Hinton, this issue; Ranzato, Huang, Boureau, & LeCun, 2007). These developments have continued to the present day, with ongoing successes at the forefront of machine speech and object recognition and many other domains. We consider these developments in more detail under *the current theoretical landscape discussion* below.

4.5. Impact on theories in particular cognitive domains

PDP has also had a strong impact on the development of specific theories in particular domains within cognitive science. The progression of these ideas, and the current state of the particular theories, are the focus of several of the remaining articles in this collection. We briefly review each here to highlight what we view as the major contributions. We note that a very broad range of researchers have contributed to these developments; only a few of these can be cited here.

4.5.1. Interactive processes in perception, language, and cognition

Some of the earliest PDP models were targeted at understanding context effects: how and why the processing of a given representational element is influenced by the concurrent processing of other elements (see McClelland et al., this issue). These ideas were explored in two computational models that have continued impact today: The interactive activation (IA) model of visual letter and word recognition (McClelland & Rumelhart, 1981)and the TRACE model of speech perception (McClelland & Elman, 1986), and these ideas have been extended to processing of words in context, using both interactive activation-like models (McRae, Spivey-Knowlton, & Tanenhaus, 1998; Spivey & Tanenhaus, 1998) and outgrowths of Elman's simple recurrent network (Tabor, Juliano, & Tanenhaus, 1997). The interactive activation and competition process embodied in these models has been adopted to address topics as wide-ranging as generalization in memory (Kumaran & McClelland, 2012; McClelland, 1981), face processing (Burton, Bruce, & Johnston, 1990), and person perception (Freeman & Ambady, 2011). McClelland et al. review the early work of Rumelhart (1977) that led to the IA model, as well as the subsequent uses of the model, and criticisms that arose about the approach, then present a new version of the model that makes explicit its connection to probabilistic Bayesian models of optimal inference and illustrates how such inference can arise as an emergent consequence of interactions among simple processing units, as Rumelhart originally envisioned.

4.5.2. Reading and language processing

Research generated by controversies surrounding early models of past-tense inflection (Rumelhart & McClelland, 1986) and single-word reading (Seidenberg & McClelland, 1989; Sejnowski & Rosenberg, 1987) has evolved considerably since these early models were proposed (Harm & Seidenberg, 2004; Rueckl, Mikolinski, Raveh, Miner, & Mars, 1997; Sibley, Kello, Plaut, & Elman, 2008). This work, and its connection to more general issues in theories of language processing, is reviewed by Seidenberg and Plaut (this issue).

4.5.3. Optimality theory

Several of the principles outlined were exemplified by Smolensky's (1986) model of cognition and language processing known as *Harmony Theory* (Smolensky & Legendre, 2005). This theory has been applied extensively to topics in sentence comprehension and phonology, and it has led to a revolutionary new theory of phonological representation known as *Optimality Theory* based on graded constraints (Prince & Smolensky, 1997). New developments in this theory are described in the paper by Smolensky, Goldrick, and Mathis (this issue).

4.5.4. Long-term memory

Current views of long-term memory have been strongly shaped by both the successes and failures of PDP models of memory. Early models (Anderson, Silverstein, Ritz, & Jones, 1977; McClelland & Rumelhart, 1985) addressed the emergence of general knowledge from experience with examples. Attempts to extend these ideas to multi-layer networks exposed the catastrophic interference problem (McCloskey & Cohen, 1989), which then led to the complementary learning-systems view of long-term memory (McClelland, McNaughton, & O'Reilly, 1995) that has continued to develop up to the present (Kumaran & McClelland, 2012; McClelland, 2013a,b; Norman & O'Reilly, 2003; O'Reilly & Rudy, 2000), as reviewed by O'Reilly et al. (this issue).

4.5.5. Semantics/conceptual knowledge

In the 1980s, similarity-based approaches to conceptual representation gave way to an alternative view under which knowledge acquisition was seen as shaped by naïve causal theories (Keil, 1989; Murphy & Medin, 1985). This research drew into focus phenomena that strongly challenged the exemplar and prototype models that characterized similarity-based approaches, but it was never wed to a comparably mechanistic account of knowledge acquisition. Rogers and McClelland (2004) were able to show, however, that many of the key phenomena emerge spontaneously from domain-general learning mechanisms in the PDP framework, which thus provides a new account of conceptual development and knowledge acquisition that does not require pre-specified domain knowledge. A rigorous mathematical analysis of the way in which these processes unfold developmentally in response to experience has recently been developed (Saxe, McClelland, & Ganguli, 2013). Insights into other aspects of semantic and conceptual development have been provided by modeling work from many other investigators (Mareschal, 2000; McMurray, Horst, & Samuelson, 2012; Westermann & Mareschal, 2012).

4.5.6. Cognitive control

A central problem for cognitive theory concerns the distinction between controlled and automatic processes. Whereas some processes seem to require slow, deliberative and conscious mental effort, others appear to unfold almost automatically, with little effort or even awareness. A central development since the publication of the PDP volumes has been the elaboration of a theory of cognitive control that treats both control and automaticity as matters of degree, in accordance with the general precepts of the framework. The origins of this work (Cohen, Dunbar, & McClelland, 1990; Phaf, van der Heijden, & Hudson, 1990) and its subsequent development—illustrating how neural networks can learn to control their behavior to conform to task goals without positing a specialized central executive—are reviewed by Botvinick and Cohen (this issue).

4.5.7. Sequential processing

Many aspects of behavior occur over time, and behavior is sensitive, not just to immediate context constraints but also to temporal context. A central contribution of PDP has been to provide a framework for understanding how temporal structure can shape learning, representation, and online processing, as exemplified in the early work of Elman reviewed above. One well-known outcome of this work has been the demonstration that such temporal sensitivity can lead to the acquisition of quite abstract implicit knowledge structures, including knowledge structures that have many of the properties of a grammar (Cleeremans & McClelland, 1991; French, Addyman, & Mareschal, 2011). Another outgrowth of this work has been its impact on current views of conscious and unconscious processing—a connection elaborated in the current issue by Cleeremans.

5. Limitations and controversies

In the years since the PDP books appeared, the framework has attracted its fair share of controversy and criticism. In some cases the controversies promoted the development of new models, leading to important new insights. In other cases, these points have led to the belief in some quarters that the framework is fundamentally flawed in important respects. We here address what we take to be the central criticisms. In general, we agree that early models (and, indeed, all existing models) do have their limitations, but this is true of models of other kinds as well. We suggest that negative conclusions that some have drawn about the validity of the core principles of PDP or the usefulness of the approach have been overstated.

5.1. Limitations and criticisms of back propagation

Earlier we noted several appealing aspects of the PDP approach to learning. Despite these features, the utility of PDP learning models for understanding human learning today remains controversial. One reason is probably the frequent use of backpropagation—the form of learning where a network's outputs are compared to an external "teaching" signal that specifies what the outputs should have been, and weights throughout the system are adjusted to reduce the discrepancy. Though backpropagation is capable, with appropriate parameterizations and training experience, of acquiring essentially any input–output mapping, it has several properties that, for some, render it an unlikely candidate model for human learning. First, standard backpropagation is fully supervised—each input is paired with a "teaching" signal that specifies the correct output. In many learning scenarios, it seems unlikely that people receive fully supervised experience. Second, backpropagation can be quite slow, requiring thousands or even millions of learning trials to master some behaviors. Third, some backprop networks suffer from catastrophic interference: Rapid learning of new material can sometimes "over-write" prior learning, leading the system to fail on previously learned material. Fourth, backpropagation learning is typically held to be biologically implausible because it requires the transmission of information "backward" from receiving units to the sending units, or even backward through time.

The criticism of backprop as biologically infeasible is, in our view, based on too literal an interpretation of the algorithm. As we stressed in discussing learning in PDP models above, the key insight has been to characterize the goal of learning in terms of optimizing an objective function. The actual implementation, we of course agree, must be consistent with the affordances of neural hardware, but the process may not literally involve propagation of error signals backward through synaptic connections. Very early on (Hinton & McClelland, 1988), it was observed that the relevant gradient information can be computed from differences between activation signals at different times, and this idea has been embedded in many successful algorithms (Hinton & Salakhutdinov, 2006; O'Reilly & Munakata, 2000).

The concern about the need for a teaching signal needs to be carefully considered, since the teacher may be nothing more than the sequence of experiences in the environment, both within and across modalities. Both auto-encoders (networks that learn to form constrained representations that allow them to reproduce their input on their outputs) and recurrent networks can be trained with backpropagation, but require only the sequence of unlabeled training data; of course, learning the relations between spellings and sounds or objects and their names does require exposure to situations in which these items occur in temporal proximity, but in general, no special teacher is required. On the other hand, a limitation of early models was their reliance on the propagation of signals all the way through many randomly initialized layers of connections between units representing inputs and others representing outputs. Much of the recent progress in machine learning discussed earlier has involved strategies for remediating this problem—for instance, training hidden layers with the additional pressure to reproduce their own inputs, allowing reliance on both labeled and unlabeled data, and the introduction of constraints that avoid reliance on idiosyncrasies and ensure robustness. These developments have led to interesting new cognitive applications (Stoianov & Zorzi, 2012).

None of this is to say that back propagation learning can address all aspects of the learning problem. The field of reinforcement learning, for example, focuses on understanding how learning systems can learn appropriate actions when there is no model of the action itself but only information about its consequences in the form of distal outcomes (Barto, 1994; Jordan & Rumelhart, 1992). A robust research program in reinforcement learning has led to the development of algorithms that adhere to general commitments of PDP and allow for the learning of complex sequential behaviors using mechanisms that bear striking resemblance to those observed in a variety of neural structures (Montague, Dayan, & Sejnowski, 1996; Suri & Schulz, 2001). Successors to these models are now helping to drive research into routine and goal-directed sequential action, as described by Botvinick and Cohen (this issue).

All current learning models still have their limitations. The development of algorithms that are directly informed by neuroscience, yet are capable of exhibiting the same flexibility and power as backpropagation while overcoming its limitations, continues as an important ongoing research agenda. We agree that much more progress is needed before we fully understand how changes in connections that support effective adaptation occurs, and this effort must continue at many levels encompassing both computational and biological considerations.

5.2. Lack of transparency

A second limitation of PDP models is the fact that they can often be opaque: Especially for consumers of the research it can be difficult to fully understand a model's behavior, and in particular, to understand why it succeeds or fails (McCloskey, 1991). Relatedly, it may be possible to show empirically that a given learning model generally shows a particular pattern of behavior but difficult to formally prove that it is guaranteed to do so or to delimit the conditions necessary for the pattern to emerge. Models are useful partly because they can help the theorist to develop new insights into the causes of otherwise puzzling patterns of data, but also because they help the theorist to communicate these insights to others. The mechanisms supporting behavior in an implemented model are fully observable and open to experimental manipulation. Thus, it is possible to work out in great detail why a model exhibits an interesting set of behaviors, which in turn can provide for clear and detailed hypotheses about why similar patterns are observed in human behavior. If, however, the causes of the model behavior remain opaque to the theorist or are not made sufficiently clear to the reader, lack of transparency can be a serious limitation.

Several lengthy debates in the literature ultimately stem from imperfect understanding of the causes of a model's behavior. In such cases, critics have rightly pointed out ways in which a particular model fails to capture certain aspects of behavioral data—then concluded that one or more of the core theoretical claims is at fault, without considering that the failure might arise instead from model attributes adopted for simplicity or as arbitrary assumptions whose importance was not initially understood. As one example, Massaro (1989) noted that predictions generated by the interactive activation model in a letter-recognition task differed systematically from observed responses, and concluded that the discrepancy arose because processing in the model was interactive—a core theoretical claim. Subsequent research has shown, however, that the real problem was the lack of intrinsic variability—another core property that was omitted from the original model (see McClelland et al., this issue). Other well-known examples include the observation of empirical shortcomings in Rumelhart and McClelland's (1986) past tense model (Pinker & Prince, 1988) and in Seidenberg and McClelland's (1989) model of single word reading (Besner, Twilley, McCann, & Seergobin, 1990). In both cases critics claimed the fundamental problem was the use of a single mechanism to process both regular and exceptional items, but subsequent research showed that many of the shortcomings depended on the models' particular input and output representations—better choices overcame the empirical shortcomings in both models (MacWhinney & Leinbach, 1991; Plaut et al., 1996), leaving the fundamental principles intact.

We believe that these and other controversies stem at least partly from a lack of transparency in the models, and thus we certainly understand the appeal of approaches in which the underlying causes of behavior seem less opaque. One approach has been to show directly how models that adhere to many of the principles of PDP can implement or underlie representations of items and structures posited at a more abstract (symbolically structured) level (Smolensky & Legendre, 2005; Smolensky et al., this issue). We also believe, however, that transparency may not be among the forces that shape the neural mechanisms underlying cognition, so that these mechanisms may not be fully penetrable at the level of explicit computational theory (this may be a general problem for science, see Morowitz, 2002). This possibility accords with the previously stated view that symbolic representations will only ever approximately characterize human cognitive abilities. For this reason, we seek alternatives to this approach.

One productive way of overcoming the challenges posed by lack of transparency may be to work toward a deeper analytic understanding of why networks behave the way that they do and how this behavior can arise from an interaction between network characteristics and the environment. To that end, we have collaborated with others to develop mathematical analyses of certain specific types of models (McClelland, 2013a,b and Movellan & McClelland, 2001 address the interactivity issue; Saxe et al., 2013 address the time course and outcome of learning in feed-forward networks like those in Fig. 1A). Further work of this kind is essential.

5.3. PDP models cannot capture abstract relational structure

In recent years, a widespread view seems to have arisen that PDP networks are unable to learn or express the abstract relational knowledge that appears to characterize human behavior in many domains (Griffiths et al., 2010; Marcus, 2001; responses to Rogers & McClelland, 2008a,b). To get a sense of the concern, consider the Farah and McClelland (1991) model discussed earlier, in which semantic representations are cast as patterns of activity across sets of semantic features—has eyes, can move, is yellow, and so on—so that properties true of the item being represented are active in its representation. Such a representation may seem insufficient because it does not express the relationship between a given property and the item with which it is associated. The relationship between cars and wheels is very different than that existing between cars and drivers. Representing both kinds of information in exactly the same way, say by activating a unit for "wheels" and another for "driver" in the semantic representation of a car, seems to miss this important distinction. Moreover, different kinds of relations support different kinds of inferences: Knowing that a Corvette is a car and that all cars have wheels supports the inference that a Corvette has wheels. In contrast, knowing that Corvettes have wheels and that all wheels are round does not support the conclusion that Corvettes are round. The "has" relation that exists between Corvettes and wheels does not support the same kind of inferences as the class-inclusion ("is a") relationship that exists between Corvettes and cars. To explain such phenomena, a theory must stipulate how knowledge of the relations between items and their properties is represented and used.

We certainly agree that relational knowledge is important for many cognitive domains and ought to be an important focus of research. The view that PDP models are constitutionally incapable of shedding light on such abilities is more puzzling. The distributed representation of relational structure has been a strong focus of research in PDP networks from quite early on. Both Hinton (1989) and Rumelhart (1990) described models of conceptual knowledge that emphasized how internal representations can be jointly shaped by knowledge of the item and the relation of interest and a recent model of analogical reasoning shows the ongoing usefulness of these ideas (Kollias & McClelland, 2013). Rumelhart's model represents different similarity structure among the same items depending upon the relation of interest, activates different sets of attributes for the same item depending upon the relation, learns similarity structure among different relation types, and can use this information productively—inferring, for instance, that items with the same basic-level name should also share similar parts but will not necessarily share similar colors (Rogers & McClelland, 2004). Several investigators have shown how knowledge of abstract grammatical relations, both in natural language and in laboratory learning studies with infants and adults, can emerge through learning in PDP networks, as a consequence of principles similar to those uncovered by Elman (Altmann & Dienes, 1999; Chang et al., 2006; McClelland, 1989; Rohde & Plaut, 1999). Likewise, the coding of abstract temporal structure in learned action sequences has been extensively investigated (see Cleeremans, this issue), and recent work with tensor-product representations has suggested one way in which abstract structural representations can be bound to particular content in fully distributed representations (Rasmussen & Eliasmith, 2011; Smolensky et al., this issue). Models that learn to form such bindings and can do so for an open-ended set of relations have also been explored extensively by several different authors (Miikkulainen, 1996; Pollack, 1990; Rohde, 2002; St. John & McClelland, 1990; St. John, 1992). Recently, many of the approaches explored in these models have been refined and integrated in machine learning models of language processing (Socher, Bauer, et al., 2013; Socher, Perelygyn, et al., 2013). While some of this work relies on more traditional methods to provide a parse tree within which neural network representations are embedded (Socher, Perelygyn, et al., 2013), another paper from this group shows that neural network learning methods can actually improve on the adequacy of the parse produced by standard methods (Socher, Bauer, et al., 2013). Given the rapid pace of current development and ongoing efforts to avoid built-in structure, it is hard to predict the extent to which pre-specified structure will ultimately be necessary.

5.4. PDP models are too flexible

The last source of controversy we will consider is the flexibility of the PDP framework. Practitioners appear to be at liberty to choose any network architecture, environment, training parameters, learning algorithm, or performance criterion they wish. There is flexibility in how inputs and outputs are represented, or in the number of units included in each layer, or in how the patterns in the environment are sampled. Different models make different choices, and it is not always clear which decisions are central to the network behavior and which peripheral. There appears to be a danger, given the array of options, that one can get these models to do anything if one just fiddles around with options until a successful combination is found. If it is possible to construct a network that will show literally any pattern of behavior, the framework may seem to be nondisconfirmable.

While we acknowledge that the framework has considerable flexibility, there are also robust and general characteristics of networks behavior that cannot so easily be overcome. The gradual nature of learning in PDP models is one such example. Here, where others have criticized the models for failure to show radical transformations they claim can occur suddenly during development (Marcus et al., 1992; Van Rijn, Van Someren, & Van Der Maas, 2003), PDP researchers have responded, not by tweaking their models to exhibit these phenomena, but by carefully re-examining the evidence and pointing out that the sudden transitions are generally not so sudden after all (McClelland & Patterson, 2002b; Schapiro & McClelland, 2009).

Humans can, however, retain the content of specific episodes far better than the types of neural network models we believe underlie experience-driven acquisition of structure-sensitive cognitive abilities. If such networks are forced to learn such information by repeated exposure without interleaving with other experiences, and if the newly learned information is structured quite differently from prior learning experiences, the new learning can catastrophically interfere with the prior knowledge, contrary to what is seen in human learners. Again, rather than tweaking the models to try to overcome such limitations, our approach has been to use such limitations to inform our understanding of the functional architecture of human memory systems. The resulting complementary learning systems theory (McClelland et al., 1995; O'Reilly et al., this issue) built on the insights of others (Marr, 1971) to address how the brain solves this problem, and is consistent with the specialized and time-limited role of the medial temporal lobes in many forms of learning and memory. Thus, the response was not simply to alter existing models to avoid catastrophic interference; instead, it lay in proposing how the computational advantages of the types of networks that gradually acquire structure sensitive cognitive abilities may be retained while simultaneously solving the catastrophic interference problem and linking these solutions to long-standing observations about the nature of different forms of long-term memory and their reliance on different brain structures.

A skeptic might reply that this example simply illustrates the problem with flexibility —the original theory and model can be altered to provide a solution to the catastrophic interference problem that maintains consistency with the core commitments of the framework. Acknowledging this, we suggest that this approach has been fruitful in providing ideas that guide experimental research as well as ongoing theory development.

6. The current theoretical landscape and the future of PDP

The current theoretical landscape differs radically from that of 1986, and we believe that, controversies and limitations notwithstanding, many shifts in the field at large reflect increasingly widespread acceptance of ideas central to the PDP framework. In retrospect one can view the PDP volumes as a distillation of several currents of thought that, in the intervening years, have shaped the discipline in pervasive ways. We here consider some of these field-wide trends, then evaluate the role of PDP as a modeling and theoretical framework in this context.

6.1. Probabilistic models of cognition

The PDP approach was partly motivated by cases in which cognition depends on graded or probabilistic constraints, and the emergence of explicitly probabilistic models of cognition in the last several years is certainly consistent with this motivation. Indeed, Rumelhart cast his initial ideas about interactive processing in perception and comprehension in explicitly probabilistic terms before adopting the less explicitly probabilistic PDP framework. Pearl's (1982) ideas about optimal Bayesian combination of both bottom-up and top–down constraints drew directly on Rumelhart's work, and Hinton and Sejnowski's (1986) Boltzmann machine presaged other explicit links between PDP models and probabilistic models of processing and learning. To a considerable degree, then, one can see PDP models as precursors of contemporary probabilistic models of cognition. Yet today, many researchers ground their research in explicitly probabilistic terms, often formulating their models as rational or computational-level models, and employing probabilistic computational methods without regard to their possible biological implementation. Often, such models address similar topics as PDP models. For example, Feldman, Griffiths, and Morgan (2009) address category effects in speech perception, a topic also addressed by the TRACE model of speech perception, and Kemp and Tenenbaum (2009) have applied Bayesian methods to the problem of learning domain structure from examples, a topic addressed by the Rumelhart model of semantic cognition (Rogers & McClelland, 2004, 2008a).

Nevertheless, PDP models differ from explicit probabilistic models in important respects noted earlier. In particular, some probabilistic models set as their goal the selection or construction of an explicit symbolic representation or other explicit structure type (Kemp & Tenenbaum, 2008, 2009), while researchers pursuing the PDP approach treat such representations as approximate characterizations of the emergent consequences of interactions between simple processing units (McClelland et al., 2010). While some may see probabilistic models as replacing or subsuming PDP models, another perspective is that the probabilistic framework and the PDP approach have overlapping but diverging aspirations. We expect both approaches to continue to evolve and to challenge each other, and possibly to benefit from a degree of competition between then.

6.2. Role of "statistical learning" widely appreciated in development

While probabilistic models have been emerging, the idea that cognitive development is strongly driven by learning about the statistical structure of the environment has also blossomed in the past 20 years. Perhaps most notable was the study by Saffran, Aslin, and Newport (1996) demonstrating that preverbal infants can rapidly learn to exploit the sequential structure of an artificial language in order to parse a continuous speech stream. This work, remarkably consistent with Elman's proposals in the early 1990s, shifted the focus of empirical research in infant development away from efforts to measure infant competencies at increasingly younger ages, and toward understanding experience-dependent mechanisms that guide cognitive change. For many researchers both within and outside the PDP tradition, the empirical mission is now to understand the nature of the information available to the infant, and the ability of the developing mind to exploit that information over time—a program that has proven extremely fruitful in advancing our understanding of language acquisition (Saffran, 2003), visuo-spatial cognition (Johnson & Johnson, 2000), causal reasoning (Cohen, Rundell, Spellman, & Cashon, 1999), and conceptual development (Sloutsky, 2010). Indeed, Clark (2013) recently argued that, across many cognitive domains, it is useful to think of the mind as an implicit prediction engine, a proposal that accords well with the PDP view of learning, representation, and processing.

6.3. Acceptance of sensitivity to both specific and general information

Another general trend has been an increasing acceptance of the view, especially within linguistics, that knowledge of language structure must reflect simultaneous sensitivity to both general and specific information. We have seen that SAID and PINT exhibit some degree of idiosyncrasy together with partial conformity to regular inflectional or spelling sound patterns, and that these idiosyncrasies tend to be shared among similar examples (c.f. KEPT, WEPT, SWEPT; BOOK, COOK, TOOK). Similar quasiregular phenomena can be found in derivational morphology (Bybee, 1985) as well as larger clausal or phrasal patterns called collocations or constructions (Goldberg, 2003). Examples like "John saw a lawyer" or "She felt the baby kick" are thought to reflect both general information about typical relationships between syntactic and semantic roles as well as more specific information captured by the particular constellation of elements; saw (but not observed or heard) a lawyer or other professional conveys "consulted," for example. There appears to be an emerging consensus that models of language knowledge must capture general and specific structure in all these different cases, and recent reliance on exemplar models (Goldinger, 1998; Pierrehumbert, 2003), similarity-based interactions between lexical items (Bybee, 1995), and construction grammars (Goldberg, 2003) all reflect attempts to capture some of these phenomena using approaches that we see as sharing similarity with (if not inspiration from) PDP approaches.

6.4. Resurgence of neural networks in machine learning

Earlier, we alluded to the recent resurgence of neural networks in machine learning. Indeed, as of this writing, the speech recognition system that allows Android phones to search the Web is a multi-layered neural network (Mohamed, Dahl, & Hinton, 2009), contemporary object classification systems for machine vision using neural networks have achieved remarkable successes (e.g. Le, 2013), and neural network models of language processing are beginning to capture the human ability to, for example, correctly interpret the sentiment expressed in complex sentences like “There are slow and repetitive parts but it has just enough spice to keep it interesting” (Socher, Perelygyn, et al., 2013). Currently, there is considerable uncertainty surrounding the question of the key factors responsible for these successes. Some important developments can be identified: (1) discovery of ways to rapidly train “deep networks”—those with at least two layers of modifiable connection weights before a classification stage—by using unlabeled data and training only one layer at a time; (2) relying on constraints on representational complexity (usually by including a pressure to allow only a few units to be active in each layer at a time, Olshausen & Field, 1996), (3) replicating learned filters and pooling outputs of neighboring filters (LeCun, Kavukcuoglu, & Farabet, 2010), (4) the use of a restricted tensor product representation to capture interactions among constituents (Socher, Bauer et al., 2013; Socher, Perelygyn, et al., 2013) and (5) integration of PDP-style learning and distributed representations within explicitly structured syntactic representations (also used in Socher, Bauer et al., 2013; Socher, Perelygyn, et al., 2013). Hinton’s paper in this special issue describes several of these developments. Another perspective, however, points to the fact that all of these key ideas were first introduced in some form in the 1980s or very soon thereafter (Ackley et al., 1985 introduced unsupervised learning; Pollack, 1990 integrated learned distributed representations into structured syntactic representations; the article on backpropagation in the PDP volumes by Rumelhart et al., 1986 introduced the replication of learned filters; Smolensky, 1990 initiated the exploration of tensor product representations; and Weigand, Rumelhart, & Huberman, 1991 explored constraints on computational complexity). The most important developments since that time may be the availability of very large corpora of training materials and very large computational engines on which the networks can be implemented. The idea has arisen that as the scale of experience and computation begins to approach the scale of experience and computation available to a young child—who sees millions of images and hears millions of words per year, and whose brain contains 10–100 billion neuron-like processing units each updating their state on a time scale of milliseconds—the full power and utility of neural networks to capture natural computation is finally beginning to become a reality, allowing artificially intelligent systems to capture more fully the capabilities of the natural intelligence present in real biological networks in the brain.

6.5. The advent of computational cognitive neuroscience

One major development in the last 25 years has been the explosive growth of computational cognitive neuroscience. The idea that computer simulations of neural mechanisms might yield insight into cognitive phenomena no longer requires, at least in most quarters, vigorous defense—there now exist whole fields, journals, and conferences dedicated to this pursuit. One consequence is the elaboration of a variety of different computationally rigorous approaches to neuroscience and cognition that capture neural information processing mechanisms at varying degrees of abstraction and complexity. These include the dynamic field theory, in which the core representational elements are fields of neurons whose activity and interactions can be expressed as a series of coupled equations (Johnson, Spencer, & Schöner, 2008); the neural engineering framework, which seeks to understand how spiking neurons might implement tensor-product approaches to symbolic representations (Eliasmith & Anderson, 2003; Rasmussen & Eliasmith, 2011); and approaches to neural representation based on ideal-observer models and probabilistic inference (Deneve, Latham, & Pouget, 1999; Knill & Pouget, 2004). Though these perspectives differ from PDP in many respects, all of these efforts share the idea that cognition emerges from interactions among populations of neurons whose function can be studied in simplified, abstract form.

6.6. Distributed representations are being taken seriously by cognitive neuroscience

Early functional brain-imaging methods were based on pre-PDP-era conceptions of mind and brain. For instance, the standard method of subtraction—in which the theorist searches for sets of anatomically contiguous voxels that are significantly more active in an experimental than a control condition—assumes that cognitive functions are localizable to anatomically discrete regions of the brain, that voxel activations can be interpreted independent of the states of other voxels, and that neighboring voxels are likely to encode the same information in the same way. Though these methods have undoubtedly shed light on the neural bases of many interesting cognitive phenomena, their widespread adoption in the 1980s and 90s led to a highly localized view of brain function in the neuro-imaging community that was very difficult to relate to the distributed, graded, and interactive view of cognition adopted by PDP.

Recently, this has begun to change, thanks to the advent of new methods for investigating distributed representations in the brain, including methods for multi-voxel pattern analysis (MVPA), where the theorist attempts to “decode” patterns of activation across large sets of voxels (Pereira, Mitchell, & Botvinick, 2009), and methods for uncovering functional and anatomical connectivity in the brain (Bullmore & Sporns, 2009). These still-recent and developing advances provide, for the first time, methodological tools that align well with the theoretical commitments of PDP. In some cases they are also revising common views of neural representation and processing in ways that are highly consistent with PDP. For instance, MVPA analyses of object representations in the ventral posterior temporal lobes have shown that voxels in the putative "fusiform face area" (FFA) can differentiate objects from houses, while voxels outside this area can differentiate faces from other kinds of objects (Haxby et al., 2001). In other words, the representation of objects, houses, and faces seem to involve a distributed code across ventral-posterior fusiform in which the FFA and neighboring regions all participate, in contrast to the alternative view that each of these stimulus categories is processed by its own dedicated and anatomically segregated cortical area.

These developments may allow the framework to realize more fully its early promise of drawing a tighter connection between cognition and neuroscience. The original PDP volumes expressed considerable optimism toward this goal, and in a few cases it is clear that the framework has helped to shape, and has been shaped by, observations from neuroscience. For instance, our current understanding of the function of the hippocampus, and how this arises from both large and small-scale aspects of its anatomy, owes much to the complementary learning systems hypothesis (O'Reilly et al., this issue). Many theories of disordered behavior following brain injury rely heavily on aspects of the PDP framework, especially in language and memory (Shallice & Cooper, 2011); and current approaches to the neuroscience of cognitive control and its disorders likewise have evolved from accounts based on PDP models (Botvinick and Cohen, this issue).

Nevertheless, in many cases, PDP models have been advanced primarily at a cognitive level, without any intended connection to underlying neural mechanisms beyond the adherence to general principles inspired by neuroscience described earlier. At the same time, cognitive neuroscience, driven largely by technological advances in brain imaging, has advanced rapidly and often without connection to the basic tenets of PDP. With the new multivariate methods and technologies for better estimating both structural and effective connectivity in the brain, this is beginning to change. In some cases, the new analysis methods capture insights derived from PDP models. For instance, representational similarity analysis—the approach in which patterns of evoked BOLD responses are interpreted by searching within cortical "spotlights" for local regions that express a stipulated similarity structure (Kriegeskorte, Mur, & Bandettini, 2008)—accords well with the observation that information in the hidden layers of PDP networks is often carried by the similarity structure of distributed representations, rather than by the mean level of activation of individual units. In other cases, observations from neuroscience are now being used to constrain the architecture of simulated neural network models, allowing theorists to develop explicit hypotheses about the functional consequences that arise from the observed anatomy. For instance, Ueno et al. (2011) described a model of single-word processing (comprehension, production, and repetition) in which the architecture of the model was constructed to conform as closely as possible to what is known about the gross anatomy and connectivity of the perisylvian language network. Within this architecture, all of the classical forms of aphasia arise from simulated damage to model analogs of the cortical regions known to produce the corresponding syndromes. Moreover, diagnostic simulations showed that performance on all three tasks suffered when the model adopted alternative internal architectures, thus providing a hypothesis about why the anatomy of the language system is structured in a particular way.

Many other examples of PDP models bridging the divide between cognitive theory and neuroscience have emerged in recent years. The well-known proposal that the anterior temporal lobes constitute a cross-modal and domain-general "hub" for semantic knowledge stemmed in part from the observation that PDP networks conforming to this architecture exhibit properties that help to resolve several puzzles in studies of concept acquisition in development (Rogers & McClelland, 2004). Our developing understanding of how functional specialization emerges within the posterior fusiform cortex—especially the putative "fusiform face area" and "visual word-form area," which reside in homologous cortex across the hemispheres—is being shaped by PDP models of learning that are tied more directly to hypotheses about the anatomy and connectivity of the early visual system (Plaut & Behrmann, 2011). The "triangle" model of word-reading, long used to understand patterns of disordered behavior, is now guiding analyses of fMRI data in participants as they read regular, irregular, and pseudo-words (Graves, Desai, Humphries, Seidenberg, & Binder, 2010). The fine-grained anatomy of the prefrontal cortex and its connectivity to the basal ganglia are being interpreted with reference to hypotheses about how PDP-like networks can possibly implement the special functions supported by these systems, that is, sustaining task-relevant information and updating it at different temporal scales as tasks are carried out (Frank, Loughry, & O'Reilly, 2001). Hypotheses about how neural systems might exploit temporal structure in learning about objects, inherited largely from Elman's work described earlier, are motivating analyses of fMRI data to uncover the brain regions where such learning has occurred (Schapiro, Rogers, Cordova, Turk-Browne, & Botvinick, 2013). Thus, we are seeing widespread efforts to employ PDP networks as tools, not only for cognitive scientists to understand healthy and disordered behaviors, but for cognitive neuroscientists to understand and interpret patterns of functional activation, and to derive hypotheses about how neuro-anatomy contributes to cognitive function.

6.7. The future role of PDP models

Given these various developments, it is probably fair to say that many of the campaigns mounted in the original PDP books, if not victorious, have at least staked out permanent settlements in the theoretical landscape. The core proposals—that cognitive structure is quasiregular, that behavior is sensitive to multiple graded constraints, that processing is graded and continuous, that representations are distributed, that cognitive development is largely driven by learning, and that it is useful to consider how cognitive functions arise from neural mechanisms—are now standard starting assumptions in many domains of research. Even if an expressed commitment to the PDP framework is avowed only by a subset of cognitive scientists and neuroscientists, many if not all of the tenets of the framework have spread very widely in the community indeed.

One might ask, then, whether the PDP framework itself remains useful for the future development of the field. For the computational cognitive neuroscientist there are now several alternative architectures to choose from, many of which adhere more concretely to detailed aspects of neural processing. For researchers in artificial intelligence, there are powerful contemporary machine learning methods that build on PDP ideas but are less tied to concerns about the details of human cognition or the possibility of implementation in the biological brain. For theoretical cognitive scientists who frame their effort at Marr's computational level, the reliance on probabilistic modeling methods (Kemp & Tenenbaum, 2008; Tenenbaum, Kemp, Griffiths, & Goodman, 2011) provides a framework for understanding cognition directly in terms of probabilistically optimal representations that satisfy many simultaneous, graded constraints without regard to their biological implementation. For the practicing experimentalist, it is even possible to conduct research that embraces many of the insights yielded by PDP without deploying the models themselves. Why then would we still need to rely on PDP models at all?

We certainly accept and celebrate the diversity of modeling approaches now in use in cognitive science and related disciplines. Models are important, not as expressions of belief, but as vehicles for exploring the implications of ideas (McClelland, 2010), and some form of modeling, whether based on PDP or otherwise, is necessary because the implications of contemporary ideas cannot always be discovered without instantiating these ideas in some form of model system. As such, as we have argued, adopting the right simplifications is essential to success in modeling, and different frameworks appropriately adopt different simplifications just as do maps made for different purposes. Each of the approaches is adapted to specific issues and applications, many of which overlap with the issues and applications addressed by the PDP framework.

Surveying the landscape of alternative approaches, we still see two very particular roles for ongoing exploration within the PDP framework. First, although many of the core PDP ideas are widely accepted in some domains, there still remain many cognitive domains where their application is seen as controversial. Even if neural network models can capture automatic processes in perception, single word reading, and even aspects of morphology, there still remains considerable skepticism about the application of such models to aspects of semantic cognition and many forms of reasoning including analogical, causal, mathematical, and syllogistic reasoning tasks. In all of these domains, we see an ongoing role for PDP models in exploring to what extent it is really necessary to make prior commitments to the use of structured representations as most existing models of these topics do.

The second role lies in connecting cognitive science to other disciplines. PDP models inhabit a kind of theoretical nexus where machine learning, cognitive and systems neuroscience, and cognitive psychology can meet and mutually constrain one another. Other neuro-computational approaches may more faithfully capture some aspects of neural processing but are generally framed at too detailed a neuro-mechanistic level, obscuring the essential properties necessary to account for the emergent cognitive phenomena. Other cognitive and machine-learning formalisms, though undoubtedly useful in many respects, either ignore or explicitly disavow any connection to neuroscience, and so remain unconstrained by any facts about the actual system that supports cognition. For theorists interested in understanding, not just how the mind works, but also how the mind arises from the brain, PDP models provide a uniquely useful tool for exploring an integrated answer to these questions.

7. Conclusion

In sum, we see an important and interesting future role for the core traditions of the PDP framework, both in extending the domain of application of the approach to higher level cognition and in continuing to serve as a bridge between disciplines and levels of analysis. We see this future work as drawing heavily on insights that have come from other approaches and other levels of analysis, and on the ever-increasing scale of computational resources and corpora of training data made available by the machine learning community. Only the future can tell just how useful ongoing explorations within the PDP framework will be. Our hope is that others will continue to participate in these explorations and to provide insights that will help guide them by pursuing complementary approaches.

Note

1. While Pinker and Prince rightly challenged the training regime that produced U-shaped performance (correct, then over-regular, then correct inflection of exceptions) in Rumelhart and McClelland's original past-tense model, Plunkett and Marchman found that the pattern of performance could be observed with a more realistic training regime. Rogers et al. observed U-shaped trends in other domains—this can often occur even if the statistics of the training environment remain stationary across all stages of development.

References

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science*, *9*, 147–169.

Albright, A., & Hayes, B. (2003). Rules vs. analogy in English past tenses: A computational/experimental study. *Cognition*, *90*(2), 119–161.

Altmann, G. T. M., & Dienes, Z. (1999). Rule learning by seven-month-old infants and neural networks. *Science*, *284*, 875.

Anderson, J. R. (1975). Computer simulation of a language-acquisition system. In R. L. Solso (Ed.), *Information processing and cognition: The loyola symposium* (pp. 295–349). Hillsdale, NJ: Lawrence Erlbaum.

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard.

Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. A., Silverstein, J. W., Ritz, S. A., & Jones, R. S. (1977). Distinctive features, categorical perception, and probability learning: Some applications of a neural model. *Psychological Review*, *84*, 413–451.

Barto, A. G. (1994). Reinforcement learning control. *Current Opinion in Neurobiology*, *4*, 888–893.

Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2007). Greedy layer-wise training of deep networks. B. Schölkopf, J. Platt & T. Hoffman (Eds.), *Advances in Neural Information Processing*, *19*(d), 153.

Besner, D., Twilley, L., McCann, R. S., & Seergobin, K. (1990). On the connection between connectionism and data: Are a few words necessary? *Psychological Review*, *97*(3), 432–446.

Botvinick, M. M., & Plaut, D. C. (2004). Doing without schema hierarchies: A recurrent connectionist approach to normal and impaired routine sequential action. *Psychological Review*, *111*(2), 395–429.

Brown, R. (1973). *A first language*. Cambridge, MA: Harvard University Press.

Bullmore, E., & Sporns, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nature Reviews Neuroscience*, *10*(3), 186–198. doi:10.1038/nrn2575.

Burton, A. M., Bruce, V., & Johnston, R. A. (1990). Understanding face recognition with an interactive activation model. *British Journal of Psychology*, *81*(3), 361–380. doi:10.1111/j.2044-8295.1990.tb02367.x.

Bybee, J. L. (1985). *Morphology: A study of the relation between meaning and form*. Amsterdam, the Netherlands: John Benjamins.

Bybee, J. L. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes*, *10*(5), 425–455.

Bybee, J. L., & Slobin, D. I. (1982). Rules and schemas in the development and use of the English past tense. *Language*, *58*(2), 265–289.

Caramazza, A. (1984). On drawing inferences about the structure of normal cognitive systems from the analysis of patterns of impaired performance: The case for single-patient studies. *Brain and Cognition*, *5*(1), 41–66. doi:10.1016/0278-2626(86)90061-8.

Carey, S. (1985). *Conceptual change in childhood*. Cambridge, MA: MIT Press.

Carey, S. (2000). The origin of concepts. *Cognition and Development*, *1*, 37–42.

Carey, S., & Spelke, E. (1994). Domain-specific knowledge and conceptual change. In L. A. H. & S. Gelman (Ed.), *Mapping the mind: Domain specificity in cognition and culture* (pp. 169–200). New York: Cambridge University Press.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, *113*(2), 234–272.

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

Clark, D. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(03), 181–204. DOI: http://dx.doi.org/10.1017/S0140525X12000477

Cleeremans, A., & McClelland, J. L. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General*, *120*, 235–253.

Cohen, J. D., Aston-Jones, G., & Gilzenrat, M. S. (2004). A systems-level perspective on attention and cognitive control. In M. I. Posner (Ed.), *Cognitive neuroscience of attention* (pp. 71–90). New York: Guilford Press.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing model of the Stroop effect. *Psychological Review*, *97*, 332–361.

Cohen, L. B., Rundell, L. J., Spellman, B. A., & Cashon, C. H. (1999). Infants' perception of causal chains. *Psychological Science*, *10*, 412–418.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A Dual Route Cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*, 204–256.

Colunga, E., & Smith, L. (2005). From the lexicon to expectations about kinds: A role for associative learning. *Psychological Review*, *112*(2), 347–382.

Davidson, K., Eng, K., & Barner, D. (2012). Does learning to count involve a semantic induction? *Cognition*, *123*(1), 162–173.

Deneve, S., Latham, P. E., & Pouget, A. (1999). Reading population codes: A neural implementation of ideal observers. *Nature Neuroscience*, *2*(8), 740–745.

Devlin, J. T., Gonnerman, L. M., Andersen, E. S., & Seidenberg, M. S. (1998). Category-specific semantic deficits in focal and widespread brain damage: A computational account. *Journal of Cognitive Neuroscience*, *10*(1), 77–94.

Dilkina, K., McClelland, J. L., & Plaut, D. C. (2008). A single-system account of semantic and lexical deficits in five semantic dementia patients. *Cognitive Neuropsychology*, *25*(2), 136–164.

Eliasmith, C., & Anderson, C. H. (2004). *Neural engineering: Computation, representation, and dynamics in neurobiological systems*. Cambridge, MA: MIT Press.

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

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*, 194–220.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development*. Cambridge, MA: MIT Press.

Evans, N., & Levinson, S. C. (2009). The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, *32*(5), 429–448; discussion 448–494.

Farah, M. J. (1990). *Visual agnosia*. Cambridge, MA: MIT Press.

Farah, M. J., & McClelland, J. L. (1991). A computational model of semantic memory impairment: Modality-specificity and emergent category-specificity. *Journal of Experimental Psychology: General*, *120*, 339–357.

Feldman, J. A., & Ballard, D. H. (1982). Connectionist models and their properties. *Cognitive Science*, *6*, 205–254.

Feldman, N. H., Griffiths, T. L., & Morgan, J. L. (2009). The influence of categories on perception: Explaining the perceptual magnet effect as optimal statistical inference. *Psychological Review*, *116*(4), 752–782.

Fodor, J. A. (1983). *Modularity of mind: An essay on faculty psychology*. Cambridge, MA: MIT Press.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*, 3–71.

Frank, M. J., Loughry, B., & O'Reilly, R. C. (2001). Interactions between frontal cortex and basal ganglia in working memory: A computational model. *Cognitive, Affective, & Behavioral Neuroscience*, *1*(2), 137–160. doi:10.3758/CABN.1.2.137.

Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. *Cognition*, *6*(4), 291–325.

Freeman, J. B., & Ambady, N. (2011). A dynamic interactive theory of person construal. *Psychological Review*, *118*(2), 247–279.

French, R. M., Addyman, C., & Mareschal, D. (2011). TRACX: A recognition-based connectionist framework for sequence segmentation and chunk extraction. *Psychological Review*, *118*(4), 614–636. doi:10.1037/a0025255.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition*, *23*, 183–209.

Gelman, R., & Williams, E. M. (1998). Enabling constraints for cognitive development and learning: A domain-specific epigenetic theory. In D. Kuhn & R. Siegler (Eds.), *Handbook of child psychology, Volume II: Cognition, perception and development* (pp. 575–630). New York: John Wiley and Sons.

Goldberg, A. E. (2003). Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences*, *7*(5), 219–224. doi:10.1016/S1364-6613(03)00080-9.

Goldinger, S. D. (1998). Echoes of echoes? An episodic theory of lexical access. *Psychological Review*, *105* (2), 251–279.

Gordon, P. (2004). Numerical cognition without words: Evidence from Amazonia. *Science*, *306*(5695), 496–499. doi:10.1126/science.1094492.

Gough, P. B. (1972). One second of reading. In J. F. Kavanagh, & I. G. Mattingly (Eds.), *Language by eye and by ear*. Cambridge, MA: MIT Press.

Graves, W. W., Desai, R., Humphries, C., Seidenberg, M. S., & Binder, J. R. (2010). Neural systems for reading aloud: A multiparametric approach. *Cerebral cortex*, *20*(8), 1799–1815. doi:10.1093/cercor/bhp245

Graves, A., Mohamed, A., & Hinton, G. (2013). Speech recognition with deep recurrent neural networks. *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 6645–6649). doi:10.1109/ICASSP.2013.6638947

Greenberg, J. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In J. Greenberg (Ed.), *Universals of language* (pp. 73–113). Cambridge, MA: MIT Press.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., Tenenbaum, J. B., & Griffiths, T. (2010). Probabilistic models of cognition: Exploring the laws of thought. *Trends in Cognitive Sciences*, *14*(7), 357–364.

Hahn, U., & Nakisa, R. C. (2000). German inflection: Single route or dual route? *Cognitive Psychology*, *41* (4), 313–360.

Harm, M. W., & Seidenberg, M. (2004). Computing the meanings of words in reading: Cooperative division of labor between visual and phonological processes. *Psychological Review*, *111*(3), 662–720.

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*, 2425–2430.

Hinton, G. E. (1989). Learning distributed representations of concepts. In R. G. M. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neurobiology* (pp. 46–61). Oxford, England: Clarendon Press.

Hinton, G. E., & McClelland, J. L. (1988). Learning representations by recirculation. In D. Z. Anderson (Ed.), *Neural Information Processing Systems* (pp. 358–366). New York: American Institute of Physics.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507. doi:10.1126/science.1127647.

Hinton, G. E., & Sejnowski, T. J. (1986). Learning and relearning in Boltzmann machines. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing* (Vol. 1, pp. 282–317). Cambridge: MIT Press.

Jacobs, R. A., & Jordan, M. I. (1992). Computational consequences of a bias toward short connections. *Journal of Cognitive Neuroscience*, *4*(4), 323–336.

Joanisse, M. F., & Seidenberg, M. S. (1999). Impairments in verb morphology after brain injury: A connectionist model. *Proceedings of the National Academy of Science*, *96*, 7592–7597.

Johnson, S. (2001). *Emergence: The connected lives of ants, brains, cities, and software*. New York: Touchstone.

Johnson, S. P., & Johnson, K. L. (2000). Early perception-action coupling: Eye movements and the development of object perception. *Infant Behavior and Development*, *23*, 461–483.

Johnson, J. S., Spencer, J. P., & Schöner, G. (2008). Moving to higher ground: The dynamic field theory and the dynamics of visual cognition. *New Ideas in Psychology*, *26*(2), 227–251. doi:10.1016/j.newideapsych.2007.07.007.

Jordan, M. I., & Rumelhart, D. E. (1992). Forward models: Supervised learning with a distal teacher. *Cognitive Science*, *16*(3), 307–354.

Karmiloff-Smith, A. (1992). *Beyond Modularity: A Developmental Perspective on Cognitive Science*. Cambridge, MA: MIT Press.

Keil, F. (1989). *Concepts, Kinds, and Cognitive Development*. Cambridge, MA: MIT Press.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, *105*(31), 10687–10692.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*(2), 20–58.

Knill, D. C., & Pouget, A. (2004). The Bayesian brain: The role of uncertainty in neural coding and computation. *Trends in Neurosciences*, *27*(12), 712–719.

Kollias, P., & McClelland, J. L. (2013). Context, cortex, and associations: A connectionist developmental approach to verbal analogies. *Frontiers in Psychology*, *4*, 857. doi:10.3389/fpsyg.2013.00857.

Kopcke, K. (1988). Schemas in German plural formation. *Lingua*, *74*, 303–335.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis—connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4. doi:10.3389/neuro.06.004.2008.

Kumaran, D., & McClelland, J. L. (2012). Generalization through the recurrent interaction of episodic memories: A model of the hippocampal system. *Psychological Review*, *119*(3), 573–616. doi:10.1037/a0028681.

Laird, J. E., Newell, A., & Rosenbloom, P. S. (1987). SOAR: An architecture for general intelligence. *Artificial Intelligence*, *33*(1), 1–64.

Lambon Ralph, M. A., Lowe, C., & Rogers, T. T. (2007). The neural basis of category-specific semantic deficits for living things: Evidence from semantic dementia, HSVE and a neural network model. *Brain*, *130*, 1127–1137.

Le, Q. V., Ranzato, M. A., Monga, R., Devin, M., Chen, K., Corrado, G. S., Dean, J. & Ng, A. Y. (2012). Building high-level features using large scale unsupervised learning. *Proceedings of the International Conference on Machine Learning*, Edinburgh, Scotland, UK. June, 2012. http://icml.cc/2012/papers/73.pdf

Le, Q. V. (2013). Building high-level features using large scale unsupervised learning. *2013 IEEE international conference on acoustics, speech and signal processing* (pp. 8595–8598). doi:10.1109/ICASSP.2013.6639343

LeCun, Y. L. Y., Kavukcuoglu, K., & Farabet, C. (2010). Convolutional networks and applications in vision. *Circuits and systems ISCAS proceedings of 2010 IEEE international symposium on circuits and systems* (pp. 253–256). doi:10.1109/ISCAS.2010.5537907

Lewes, G. H. (1879). *Problems of life and mind*. Vol. 1. Boston, MA: Houghton, Osgood and Company.

MacWhinney, B., & Leinbach, J. (1991). Implementations are not conceptualizations: Revising the verb learning model. *Cognition*, *40*, 121–153.

Marcus, G. F. (2001). *The algebraic mind*. Cambridge, MA: MIT Press.

Marcus, G. F., Brinkmann, U., Clahsen, H., Wiese, R., & Pinker, S. (1995). German inflection: The exception that proves the rule. *Cognitive Psychology*, *29*(3), 189–256.

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, J. T., & Xu, F. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, *57*(4), 1–165.

Mareschal, D. (2000). Infant object knowledge: Current trends and controversies. *Trends in Cognitive Science*, *4*, 408–416.

Mareschal, D., French, R. M., & Quinn, P. C. (2000). A connectionist account of asymmetric category learning in early infancy. *Developmental Psychology*, *36*(5), 635–645.

Marr, D. (1971). Simple memory: A theory for archicortex. *The Philosophical Transactions of the Royal Society of London*, *262*(Series B), 23–81.

Marr, D. (1982). *Vision*. New York: Freeman.

Massaro, D. W. (1989). Testing between the TRACE model and the fuzzy logical model of speech perception. *Cognitive Psychology*, *21*, 398–421.

Mayor, J., & Plunkett, K. (2010). A neurocomputational account of taxonomic responding and fast mapping in early word learning. *Psychological Review*, *117*(1), 1–31. doi:10.1037/a0018130.

McClelland, J. L. (1981). Retrieving general and specific information from stored knowledge of specifics. *Proceedings of the Third Annual Conference of the Cognitive Science Society* (pp. 170–172). Berkeley, CA: Cognitive Science Society.

McClelland, J. L. (1989). Parallel distributed processing and role assignment constraints. In Y. Wilks (Ed.), *Theoretical issues in natural language processing* (pp. 78–85). Hillsdale, NJ: Erlbaum.

McClelland, J. L. (2010). Emergence in cognitive science. *Topics in Cognitive Science*, *2*(4), 751–770. doi:10.1111/j.1756-8765.2010.01116.x.

McClelland, J. L. (2013a). Integrating probabilistic models of perception and interactive neural networks: A historical and tutorial review. *Frontiers in Psychology*, doi:10.3389/fpsyg.2013.00503.

McClelland, J. L. (2013b). Incorporating rapid neocortical learning of new schema-consistent information into complementary learning systems theory. *Journal of Experimental Psychology: General*, *142*(4), 1190–1210.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences*, *14*(8), 348–356.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1–86.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*, 419–457.

McClelland, J. L., & Patterson, K. (2002a). "Words or Rules" cannot exploit the regularity in exceptions. *Trends in Cognitive Science*, *6*, 464–465.

McClelland, J. L., & Patterson, K. (2002b). Rules or connections in past-tense inflections: What does the evidence rule out? *Trends in Cognitive Science*, *6*, 465–472.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review*, *88*, 375–407.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General*, *114*, 159–188.

McClelland, J. L., Rumelhart, D. E., & Hinton, G. E. (1986). The appeal of parallel distributed processing. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1, pp. 3–44). Cambridge, MA: MIT Press.

McClelland, J. L., & St. John, M. F. & Taraban, R. (1989). Sentence comprehension: A parallel distibuted processing approach. *Language and Cognitive Processes*, *4*, 287–335.

McCloskey, M. (1991). Networks and theories: The place of connectionism in cognitive science. *Psychological Science*, *2*(6), 387–395.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.), *The psychology of learning and motivation*. Vol. 24 (pp. 109–165). New York: Academic Press.

McMurray, B., Horst, J. S., & Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review*, *119*(4), 831–877.

McRae, K., & Cree, G. (2002). Factors underlying category-specific deficits. In G. Humphreys (Ed.), *Category specificity in brain and mind*. Hove, UK: Psychology Press.

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language*, *38*(3), 283–312. doi:10.1006/jmla.1997.2543.

Miikkulainen, R. (1996). Subsymbolic case-role analysis of sentences with embedded clauses. *Cognitive Science*, *20*(1), 47–74.

Minsky, M. (1974). A framework for representing knowledge. In P. Winston (Ed.), *The psychology of computer vision* (pp. 211–277). New York, NY: McGraw-Hill. doi:10.2307/3680823

Minsky, M., & Papert, S. (1969). *Perceptrons: An introduction to computational geometry*. Cambridge, MA: MIT Press.

Mohamed, A., Dahl, G., & Hinton, G. (2009). Deep belief networks for phone recognition. *Science*, *4*(5), 1–9. doi:10.4249/scholarpedia.5947.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *Journal of Neuroscience*, *16*, 1936–1947.

Morowitz, H. J. (2002). *The emergence of everything: How the world became complex*. New York: Oxford University Press.

Movellan, J. R., & McClelland, J. L. (2001). The Morton-Massaro law of information integration: Implications for models of perception. *Psychological Review*, *108*, 113–148.

Munakata, Y., & McClelland, J. L. (2003). Connectionist models of development. *Developmental Science*, *6* (4), 413–429.

Munakata, Y., McClelland, J. L., Johnson, M. H., & Siegler, R. S. (1997). Rethinking infant knowledge: Toward an adaptive process account of successes and failures in object permanence tasks. *Psychological Review*, *104*, 686–713.

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

Neisser, U. (1967). *Cognitive Psychology*. New York: Appleton-Century-Crofts.

Newell, A. (1990). *Unified Theories of Cognition*. Cambridge, MA: MIT Press.

Newell, Allen., & Simon, H. A. (1961). Computer simulation of human thinking. *Science*, *134*(3495), 2011–2017.

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary learning-systems approach. *Psychological Review*, *110*(4), 611–646.

Oden, G. C. (1978). Semantic constraints and judged preference for interpretations of ambiguous sentences. *Memory & Cognition*, *6*(1), 26–37. doi:10.3758/BF03197425.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive-field properties by learning a sparse code for natural images. *Nature*, *381*, 607–609.

O'Reilly, R. C., Braver, T. S., & Cohen, J. D. (1999). A biologically-based computational model of working memory. In A. Miyake & P. Shah (Ed.), *Models of working memory: Mechanisms of active maintenance and executive control* (pp. 375–411). New York: Cambridge University Press.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience*. Cambridge, MA: MIT Press.

O'Reilly, R. C., & Rudy, J. W. (2000). Computational principles of learning in the neocortex and hippocampus. *Hippocampus*, *10*(4), 389–397.

Patterson, K., Lambon Ralph, M. A., Jefferies, E., Woollams, A., Jones, R., Hodges, J. R., & Rogers, T. T. (2006). "Pre-semantic" cognition in semantic dementia: Six deficits in search of an explanation. *Journal of Cognitive Neuroscience*, *18*(2), 169–183.

Pearl, J. (1982). Reverend Bayes on inference engines: A distributed hierarchical approach. In D. L. Waltz (Ed.), *Proceedings of the AAAI national conference on AI* (pp. 133–136).

Pereira, F., Mitchell, T., & Botvinick, M. M. (2009). Machine learning classifiers and fMRI: A tutorial overview. *NeuroImage*, *45*(1), S199–S209.

Perry, C., Ziegler, J. C., & Zorzi, M. (2010). Beyond single syllables: large-scale modeling of reading aloud with the Connectionist Dual Process (CDP++) model. *Cognitive Psychology*, *61*(2), 106–151.

Phaf, R. H., van der Heijden, A. H. C., & Hudson, P. T. W. (1990). SLAM: A connectionist model for attention in visual selection tasks. *Cognitive Psychology*, *22*, 273–341.

Piaget, J. (1970). *The child's conception of movement and speed*. New York: Ballantine.

Piaget, J. (2008). Intellectual evolution from adolescence to adulthood. *Human Development*, *51*(1), 40–47. doi:10.1159/000112531.

Pierrehumbert, J. B. (2003). Phonetic diversity, statistical learning, and acquisition of phonology. *Language and Speech*, *46*(Pt 2–3), 115–154.

Pinker, S. (1991). Rules of language. *Science*, *253*, 530.

Pinker, S. (1999). *Words and Rules*. New York: Basic.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a Parallel Distributed Processing model of language acquisition. *Cognition*, *28*, 73–193.

Plaut, D. C. (2002). Graded modality-specific specialisation in semantics: A computational account of optic aphasia. *Cognitive Neuropsychology*, *19*(7), 603–639.

Plaut, D. C., & Behrmann, M. (2011). Complementary neural representations for faces and words: A computational exploration. *Cognitive Neuropsychology*, *28*(3–4), 251–275. doi:10.1080/02643294.2011.609812.

Plaut, D. C., & Booth, J. R. (2000). Individual and developmental differences in semantic priming: Empirical and computational support for a single-mechanism account of lexical processing. *Psychological Review*, *107*(4), 786–823.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, *103*, 56–115.

Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology*, *10*(5), 377–500.

Plunkett, K., & Marchman, V. A. (1991). U-Shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition. *Cognition*, *38*, 43–102.

Plunkett, K., & Marchman, V. A. (1996). Learning from a connectionist model of the acquisition of the English past tense. *Cognition*, *61*(3), 299–308.

Plunkett, K., & Sinha, C. (1992). Connectionism and developmental theory. *British Journal of Developmental Psychology*, *10*(3), 209–254.

Pollack, J. B. (1990). Recursive distributed representations. *Artificial Intelligence*, *46*(1–2), 77–105. doi:10.1016/0004-3702(90)90005-K.

Prince, A., & Smolensky, P. (1997). Optimality: From neural networks to universal grammar. *Science*, *275* (5306), 1604–1610.

Ranzato, M., Huang, F. J., Boureau, Y.-L., & LeCun, Y. (2007). Unsupervised learning of invariant feature hierarchies with applications to object recognition. *2007 IEEE conference on computer vision and pattern recognition* (pp. 1–8). doi:10.1109/CVPR.2007.383157

Rasmussen, D., & Eliasmith, C. (2011). A neural model of rule generation in inductive reasoning. In R. Cattrambone & S. Ohlsson (Eds.), *Topics in Cognitive Science*, *3*(1), 140–153. doi:10.1111/j.1756-8765.2010.01127.x

Reicher, G. M. (1969). Perceptual recognition as a function of meaningfulness of stimulus material. *Journal of Experimental Psychology*, *81*, 275–280.

Rogers, T. T., Lambon Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., & Patterson, K. (2004). The structure and deterioration of semantic memory: A computational and neuropsychological investigation. *Psychological Review*, *111*(1), 205–235.

Rogers, T. T., Lambon Ralph, M. A., Hodges, J. R., & Patterson, K. (2003). Object recognition under semantic impairment: The effects of conceptual regularities on perceptual decisions. *Language and Cognitive Processes*, *18*(5/6), 625–662.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. Cambridge, MA: MIT Press.

Rogers, T. T., & McClelland, J. L. (2008a). A simple model from a powerful framework that spans levels of analysis. *Behavioral and Brain Sciences*, *31*, 729–749.

Rogers, T. T., & McClelland, J. L. (2008b). Précis of semantic cognition: A parallel distributed processing approach. *Behavioral and Brain Sciences*, *31*(06), 689. doi:10.1017/S0140525X0800589X.

Rogers, T. T., Rakison, D. H., & McClelland, J. L. (2004). U-shaped curves in development: A PDP approach. *Journal of Cognition and Development*, *5*(1), 137–145. doi:10.1207/s15327647jcd0501_14.

Rohde, D. L. (2002). A connectionist model of sentence comprehension and production. Ph.D. Thesis. Pittsburgh, PA: Computer Science Department, Carnegie Mellon University.

Rohde, D. L., & Plaut, D. C. (1999). Simple recurrent networks can distinguish non-occurring from ungrammatical sentences given appropriate task structure: Reply to Marcus. *Cognition*, *73*(3), 297–300.

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408.

Rueckl, J. G., Mikolinski, M., Raveh, M., Miner, C. S., & Mars, F. (1997). Morphological priming, fragment completion, and connectionist networks. *Journal of Memory and Language*, *36*(3), 382–405.

Rumelhart, D. E. (1977). Toward an interactive model of reading. In S. Dornic (Ed.), *Attention and performance VI*. Hillsdale, NJ: Erlbaum.

Rumelhart, D. E. (1989). The architecture of mind: A connectionist approach. *Foundations of Cognitive Science*, *1*, 133–159.

Rumelhart, D. E. (1990). Brain style computation: Learning and generalization. In S. F. Zornetzer, J. L. Ddavis, & C. Lau (Eds.), *An Introduction to neural and electronic networks* (pp. 405–420). San Diego, CA: Academic Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart J. L. McClelland, & the PDP Research Group (Eds.), *Parallel distributed* *processing: Explorations in the microstructure of cognition* (Vol. 1, pp. 318–362). Cambridge, MA: MIT Press.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2, pp. 216–271). Cambridge, MA: MIT Press.

Rumelhart, D. E., & McClelland, J. L., & the PDP Research Group (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume I: foundations & volume II: Psychological and biological models*. Cambridge, MA: MIT Press.

Rumelhart, D. E., & Ortony, A. (1976). The representation of knowledge in memory. In W. E. Montague, R. Anderson, & R. J. Sprio (Eds.), *Schooling and the acquisition of knowledge*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Schemata and sequential thought processes in PDP Models. In J. L. McClelland & D. E. Rumelhart, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 2, pp. 7–57). Cambridge, MA: MIT Press.

Saffran, J. R. (2003). Statistical language learning: Mechanisms and constraints. *Current Directions in Psychological Science*, *12*(4), 110–114. doi:10.1111/1467-8721.01243.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-olds. *Science*, *274* (5294), 1926–1928.

Salakhutdinov, R., & Hinton, G. (2009). Deep Boltzmann machines. *Artificial Intelligence*, *5*(2), 448–455.

Samuel, A. L. (1959). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development*, *3*(3), 210–229. doi:10.1147/rd.441.0206.

Saxe, A., McClelland, J. L., & Ganguli, S. (2013). Learning hierarchical categories in deep networks. *Proceedings of the 35th annual meeting of the Cognitive Science Society* (pp. 1271–1276). Austin, TX: Cognitive Science Society.

Schank, R. C. (1981). Language and memory. In D. A. Norman (Ed.), *Perspectives on cognitive science* (pp. 105–146). Hillsdale, NJ: Lawrence Erlbaum Associates.

Schank, R. C., & Abelson, R. P. (1977). Scripts, plans, goals and understanding. In A. Flammer & W. Kintsch (Eds.), *Representation* (Vol. Discourse, p. 272). Hillsdale, NJ: Lawrence Erlbaum Associates.

Schapiro, A. C., & McClelland, J. L. (2009). A connectionist model of a continuous developmental transition in the balance scale task. *Cognition*, *110*(3), 395–411.

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. M. (2013). Neural representations of events arise from temporal “community” structure. *Nature Neuroscience*, *16*, 486–492.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*, 523–568.

Sejnowski, T. J., & Rosenberg, C. R. (1987). Parallel networks that learn to pronounce English text. *Complex Systems*, *1*, 145–168.

Shallice, T. (1988). *From neuropsychology to mental structure*. Oxford, UK: Oxford University Press.

Shallice, T., & Cooper, R. P. (2011). *The organization of mind*. Oxford, UK: Oxford University Press.

Shirai, Y., & Andersen, R. W. (1995). The acquisition of tense-aspect morphology: A prototype account. *Language*, *71*(4), 743–762.

Sibley, D. E., Kello, C. T., Plaut, D. C., & Elman, J. L. (2008). Large-scale modeling of wordform learning and representation. *Cognitive Science*, *32*(4), 741–754. doi:10.1080/03640210802066964.

Siegler, R. S. (1976). Three aspects of cognitive development. *Cognitive Psychology*, *8*(4), 481–520.

Siegler, R. S. (1981). Developmental sequences within and between concepts. *Monographs of the Society for Research in Child Development*, *46*(2), 1–84. doi:10.2307/1165995.

Siegler, R. S., & Alibali, M. W. (2005). *Childrens thinking*. Upper Saddle River, NJ: Prentice Hall.

Siegler, R. S., & Chen, Z. (1998). Developmental differences in rule learning: A microgenetic analysis. *Cognitive Psychology*, *36*(3), 273–310.

Sloutsky, V. M. (2010). From perceptual categories to concepts: What develops? *Cognitive Science*, *34*(7), 1244–1286. doi:10.1111/j.1551-6709.2010.01129.x.

Smolensky, P. (1986). Information processing in dynamical systems: Foundations of harmony theory. In J. L. Mcclelland, D. E. Rumelhart, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition* (Vol. 1, pp. 194–281). Cambridge, MA: MIT Press.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, *46*(1–2), 159–216. doi:10.1016/0004-3702(90)90007-M.

Smolensky, P., & Legendre, G. (2005). *The harmonic mind*. Cambridge, MA: MIT Press.

Socher, R., Bauer, J., Manning, C. D., & Ng, A. Y. (2013). Parsing with compositional vector grammars. *Proceedings of the 51st annual meeting of the Association for Computational Linguistics*(Volume 1: Long Papers) (pp. 455–465). Sofia, Bulgaria: Association for Computational Linguistics.

Socher, R., Perelygin, A., Wu, J. Y., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. *Proceedings of the 2013 conference on empirical methods in natural language processing* (pp. 1631–1642). Seattle, WA: Association for Computational Linguistics.

Spelke, E. S., Breinlinger, K., Macomber, J., & Jacobson, K. (1992). Origins of knowledge. *Psychological Review*, *99*(4), 605–632.

Spencer, J., Thomas, M. S. C., & McClelland, J. L. (2009). *Toward a unified theory of development: Connectionism and dynamic systems theory re-considered*. Oxford, UK: Oxford University Press.

Spivey, M. J. (2008). *Continuity of Mind*. Oxford, UK: Oxford University Press.

Spivey, M. J., & Tanenhaus, M. K. (1998). Syntactic ambiguity resolution in discourse: Modeling the effects of referential context and lexical frequency. *Journal of Experimental Psychology: Learning Memory and Cognition*, *24*(6), 1521–1543.

St. John, M. F. (1992). The story gestalt: A model of knowledge-intensive processes in text comprehension. *Cognitive Science*, *16*, 271–306.

St. John, M. F., McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence*, *46*, 217–257.

Sternberg, S. (1969). The discovery of processing stages: Extensions of Donders' method. *Acta Psychologica*, *30*(1), 276–315.

Stoianov, I., & Zorzi, M. (2012). Emergence of a "visual number sense" in hierarchical generative models. *Nature Neuroscience*, *15*(2), 194–196. doi:10.1038/nn.2996.

Suri, R. E., & Schulz, W. (2001). Temporal difference model reproduces anticipatory neural activity. *Neural Computation*, *13*, 841–862.

Tabor, W., Juliano, C., & Tanenhaus, M. (1997). Parsing in a dynamical system: An attractor-based account of the interaction of lexical and structural constraints in sentence processing. *Language and Cognitive Processes*, *12*(2), 211–271. doi:10.1080/016909697386853.

Taraban, R., & McClelland, J. L. (1988). Constituent attachment and thematic role assignment in sentence processing: Influences of content-based expectations. *Journal of Memory and Language*, *27*, 597–632.

Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences*, *10*(7), 309–318.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science* (New York, N.Y.), *331*(6022), 1279–1285. doi:10.1126/science.1192788.

Thelen, E., & Smith, L. B. (1996). *A dynamic systems approach to the development of cognition and action*. Cambridge, MA: MIT Press.

Thomas, M. S. C., & Karmiloff-Smith, A. (2003). Modeling language acquisition in atypical phenotypes. *Psychological Review*, *110*(4), 647–682.

Tyler, L., Moss, H. E., Durrant-Peatfield, M. R., & Levy, J. P. (2000). Conceptual structure and the structure of concepts: A distributed account of category-specific deficits. *Brain and Language*, *75*(2), 195–231.

Ueno, T., Saito, S., Rogers, T. T., & Lambon Ralph, M. A. (2011). Lichtheim 2: Synthesizing aphasia and the neural basis of language in a neurocomputational model of the dual dorsal-ventral language pathways. *Neuron*, *72*(2), 385–396. doi:10.1016/j.neuron.2011.09.013.

Van Rijn, H., Van Someren, M., & Van Der Maas, H. (2003). Modeling developmental transitions on the balance scale task. *Cognitive Science*, *27*(2), 227–257. doi:10.1207/s15516709cog2702_4.

Weigand, A. S., Rumelhart, D. E., & Huberman, B. A. (1991). Generalization by weight elimination applied to currency exchange rate prediction. *Advances in Neural Information Processing Systems 3* (pp. 875–882). Menlo Park, CA: Morgan Kaufmann.

Wellman, H. M., & Gelman, S. A. (1997). Knowledge acquisition in foundational domains. In D. Kuhn & R. Siegler (Eds.), *Handbook of child psychology*, Volume 2: *Cognition, perception and development* (pp. 523–573). New York: John Wiley and Sons.

Westermann, G., & Mareschal, D. (2012). Mechanisms of developmental change in infant categorization. *Cognitive Development*, *27*(4), 367–382. doi:10.1016/j.cogdev.2012.08.004.

Wheeler, D. D. (1970). Processes in word recognition. *Cognitive Psychology*, *1*, 59–85.

Wilkening, F., & Anderson, N. H. (1980). Comparison of two rule assessment methodologies for studying cognitive development. Center for Human Information Processing Report (Vol. 94, p. 44 pp.).

Woods, W. A. (1970). Transition network grammars for natural language analysis. B. Grosz, K. Jones, & B. Webber, (Eds.), *Communications of the ACM*, *13*(10), 591–606. doi:10.1145/355598.362773

Woollams, A. M., Lambon Ralph, M. A., Plaut, D. C., & Nagaraja, S. S. (2007). SD-squared: On the association between semantic dementia and surface dyslexia. *Psychological Review*, *114*(2), 316–339.