

Opinion

Parallel Distributed Processing Theory in the Age of Deep Networks

Jeffrey S. Bowers^{1,*}

Parallel distributed processing (PDP) models in psychology are the precursors of deep networks used in computer science. However, only PDP models are associated with two core psychological claims, namely that all knowledge is coded in a distributed format and cognition is mediated by non-symbolic computations. These claims have long been debated in cognitive science, and recent work with deep networks speaks to this debate. Specifically, single-unit recordings show that deep networks learn units that respond selectively to meaningful categories, and researchers are finding that deep networks need to be supplemented with symbolic systems to perform some tasks. Given the close links between PDP and deep networks, it is surprising that research with deep networks is challenging PDP theory.

PDP and Deep Neural Networks

Parallel distributed processing (PDP, see [Glossary](#)) theories of cognition [1,2] have had a profound influence in psychology and, recently, in computer science. With regards to psychology, PDP theories are associated with a host of fundamental claims, but here I focus on two, namely that knowledge is coded in a distributed rather than a localist format, and that computations are performed in a non-symbolic rather than symbolic manner ([Box 1](#)). As detailed below, these claims are currently the prominent view in both psychology and neuroscience, and challenge many classic theories in psychology, linguistics, and artificial intelligence.

With regards to computer science, PDP models are the precursor to recent **deep neural networks** that have achieved state-of-the-art performance across a range of tasks, including speech [3,4] and object [5,6] recognition. This in turn has led to billions of dollars in investment in developing deep networks by Google, Facebook, Baidu, among other technology companies. Strikingly, these networks are in many ways similar to PDP models. Indeed, the most common way to train deep networks is through the **back-propagation algorithm** developed for PDP models in the 1980s. Perhaps the two most important differences between the models of the past and current models is that today we have more powerful computers and **graphics processing units** (GPUs) that have sped up simulations by orders of magnitudes, and we have vastly larger datasets of labeled data for training networks. This, together with the introduction of more efficient **activation functions**, has made it possible to train networks with many layers, millions of units, and billions of connections (a recent model included over 1000 hidden layers [7]). Note that it is the many layers of these models that led to the term ‘deep’ networks, whereas earlier PDP networks tended to include only a few layers of units.

Given the similarities between PDP and deep networks, it might be expected that the successes of deep networks would lend general support to PDP theories of cognition. Indeed, the most common criticism of PDP models is that they are not sufficiently powerful to explain

Trends

Deep neural networks recently developed in computer science are similar to the PDP models developed in psychology in the 1980s.

The advent of more powerful computers and larger datasets has allowed modelers to train networks much more extensively. This has led to networks that solve real-world problems, including state-of-the-art speech and object recognition.

It is still a mystery why deep networks work as well as they do, which has led to a recent explosion of single-unit recording studies of deep networks. These studies have revealed highly selective units, similar to the neurons observed in hippocampus and cortex.

Based on some limitations of deep networks, theorists have started to develop hybrid networks that include symbolic representations and computations.

The fact that deep networks learn selective codes and need symbols to solve some tasks challenges two fundamental claims of PDP theory.

¹School of Experimental Psychology, University of Bristol, 12a Priory Road, Bristol, BS8 1TU, UK

*Correspondence: j.bowers@bristol.ac.uk (J.S. Bowers).

Box 1. Two Core Theoretical Debates

It is important to distinguish between the localist/distributed and the symbolic/non-symbolic debates. The localist/distributed debate concerns the interpretability of individual units in an artificial or real neural network. Units in an artificial network are analogous to neurons in brains in that they both respond to inputs (e.g., firing rate of a neuron) and connect to other units (neurons). The key feature of a localist unit is that it is most active to one meaningful category. For instance, in the interactive activation model [61], each word unit responds most strongly to a specific word, and, as a consequence, it is possible to interpret the output of single units (if unit X is active beyond some threshold, the model has identified the word DOG). By contrast, a representation is distributed if each unit responds to multiple categories to the same degree, and as a consequence the pattern of activation over a collection of units is necessary to uniquely categorize an input. With this view it is not possible to determine what the model has identified by observing the state of single units.

The symbolic/non-symbolic debate concerns how neural systems compute, and this entails a different claim about how knowledge is coded. A key feature of symbolic systems is that words, objects, concepts, etc. are represented in long-term memory in a format that support 'compositionality' such that complex representations are composed from simpler parts that are context-independent [9]. For example, in a symbolic model of word identification, words are coded from letters that maintain their identity in different contexts (e.g., the words DOG and GOD share the same set of letter representations despite the fact that the letters D and G occur in different positions). Crucially, symbolic networks need methods to compute with context-independent representations to dynamically assign items a role (in this case, assign letters a position), a process also called variable binding. This ensures that DOG and GOD are similar to one another given that they share the same set of letter units, but differ by virtue of the way the letters are assigned a position. This dynamic binding requires additional circuits [57] or additional computational mechanisms [10] compared to non-symbolic models that compute on context-dependent representations where the binding are coded in long-term memory (e.g., D-in-position-1 and D-in-position-3). Importantly, the use of context-dependent representations obscures the similarity of items (DOG and GOD only share the O-in-position-2 letter code). It is often claimed that symbolic models support more widespread generalization, specifically, generalization 'outside the training set' [11].

human intelligence given their commitment to distributed representations [8] and non-symbolic computations [9–11]. Accordingly, the ability of deep networks to solve some complex tasks, sometimes at super-human levels, might appear to undermine this critique.

I argue exactly the opposite and highlight how current work with deep networks is providing the most compelling demonstration to date that PDP theories of human cognition are fundamentally flawed. Two findings in particular pose a challenge for PDP theory, namely that deep networks learn highly selective representations under a range of conditions [12], and deep networks fail in solving precisely those tasks that the proponents of symbolic systems predicted all along [13]. Similar findings have been found with PDP networks, both with regards to their learned representations [8, 14, 15] and their computational limitations [10, 11]. However, given the level of attention directed to deep networks, the deep learning results may have more traction in changing minds in psychology and neuroscience.

Localist Versus Distributed Coding

One of the common arguments put forward in support of PDP theories in psychology is that distributed representations are more biologically plausible than localist representations. Indeed, localist models in psychology are often rejected on the basis that **grandmother cells** are untenable [16]. On the grandmother cell hypothesis, high-level categories (e.g., familiar words, objects, or faces) are identified when a single neuron fires beyond some threshold. However, grandmother cells in the neuroscience literature and **localist representations** in the psychology literature are often defined differently [17], and accordingly rejecting grandmother cells has little or no bearing on the biological plausibility of localist models. In fact, there is now compelling evidence that some single neurons in the hippocampus and the cortex respond to familiar high-level information in a highly selective manner [18], and single-unit recordings in localist models in psychology are consistent with a range of single-cell recording studies carried out on brains [19, 20]. When grandmother cells are defined as localist representations (to make the

Glossary

Activation function: the activation function of a unit determines the output of that unit given its inputs. Various activation functions have been employed across different networks, including sigmoidal, Gaussian, and rectifier functions.

Back-propagation algorithm: a method for training PDP and deep neural networks. It is a 'supervised' form of learning because the model is provided the correct output for each input. The algorithm adjusts the weights between units across all layers of a network such that an input does a better job in producing the correct output on later trials.

Convolutional networks: a deep network in which units in a convolutional layer are connected to a subset of spatially contiguous units in the preceding layer. Convolutional networks are faster to train than standard networks because they include fewer connections.

Deep neural networks: any network that includes multiple hidden layers.

Gaussian units: units that use a non-linear 'bell-curved' Gaussian activation function in which the output of unit changes non-monotonically with increasing inputs. Value units employ a specific form of the Gaussian function.

Grandmother cell: a hypothetical neuron that codes for one meaningful category (e.g., an image of a specific person). It is a pejorative term intended to ridicule this hypothesis, and as a consequence there have been few attempts to provide a formal definition. On one view, grandmother cells only fire in response to a single specific stimulus, with separate neurons being devoted to each possible perceptual experience. On another view, grandmother cells are the equivalent of localist representations in psychological models. The former hypothesis is clearly false, the latter view is more plausible.

Graphics processing unit (GPU): an electronic circuit that is well suited for the matrix/vector math involved in training deep neural networks. The use of GPUs started in 2009 was estimated to speed up network training by approximately 100-fold.

grandmother cell hypothesis a well-specified and serious hypothesis rather than the straw-man hypothesis that it often is), then grandmother cells are biologically plausible [17,18].

However, what is less well known, and the point I want to emphasize here, is that both PDP models and deep networks often learn localist representations.

Localist Representations in PDP Models

Within psychology, Berkeley *et al.* [14] were the first to carry out single-unit recordings on PDP networks to explore the conditions under which networks learn localist versus distributed representations for high-level information. After training their models on various complex input–output mappings, they recorded the response of each hidden unit to a range of inputs. They then displayed the results using a scatter plot for each unit. The response of each unit to a specific input was coded with a point along the *x* axis (ranging from 0 to 1), with values on the *y* axis being arbitrary (to prevent overlapping responses from different inputs; illustrated in Figure 1A). The key finding was that the networks learned some localist representations when they included **Gaussian units** (that the authors called ‘value units’). By contrast, they failed to observe localist codes when their models included **sigmoidal units** that are typically used in PDP networks. This highlights that distributed representations are not an intrinsic property of PDP networks, but are instead associated with specific implementations of PDP networks.

More recently, we adapted these scatter plots to explore the conditions in which recurrent PDP models of short-term memory (STM) learn localist codes [8,15]. The models used sigmoidal activation functions and were highly similar to those developed by Botvinick and Plaut [21]. We found that the networks learned distributed representations when they were trained to recall single items, but learned localist representations when trained to recall sequences of items. That is, we found that PDP networks learned distributed codes when the models were trained to activate one item at a time in STM, and learned localist codes when trained to coactivate multiple items in STM (Figure 1B). We argued that learned distributed representations were unable to overcome the **superposition catastrophe** [22] in the later condition, and that the models were therefore forced to learn localist codes to succeed.

This computational explanation for the emergence of localist coding in our simulations complements an earlier analysis of Marr [23]. In the same way as Marr argued that long-term memory is coded in a highly selective manner in the hippocampus to encode new memories quickly without forgetting pre-existing memories (solving the so-called stability–plasticity dilemma [24], otherwise known as catastrophic interference [25]), we argued that long-term knowledge in the cortex is coded in a selective manner to support STM (solving the superposition catastrophe). More generally, Plaut and McClelland [26] argue that PDP networks ‘discover representations that are effective in solving tasks . . .’ and this ‘provides more insight into why cognitive and neural systems are organized the way they are’ (p. 489). Adopting this logic, the conclusion must be that there are computational advantages of localist codes in some conditions, and the findings may help to explain why some neurons in cortex respond in such a highly selective manner.

Localist Representations in Deep Networks

In contrast to the handful of single-unit recording studies carried out on PDP models over the past 30 years, there has been an explosion of single-unit studies carried out on deep networks. The striking finding across dozens of studies is that the networks learn highly selective representations for familiar categories across a range of network architectures and tasks ([17] for a short review of single-unit recording studies in PDP and deep networks). Importantly,

Localist representation: a localist representation responds most strongly to one familiar meaningful category such as a word, object, or person. Although localist representations encode one and only one thing, they do respond to other related items. For example, a localist representation of the word DOG fires most strongly to the word DOG, but also will fire (below some threshold) to related words such as LOG or FOG.

Parallel distributed processing (PDP): a form of artificial neural network developed in the 1980s that was associated with a host of psychological and neuroscience claims. Two central claims are that information is coded in a distributed format, and computations are symbolic.

Sigmoidal units: units that use a non-linear S-shaped sigmoidal activation function in which the output of a unit increases monotonically from 0 to 1 with increasing inputs. These are most commonly used in PDP models.

Superposition catastrophe: a hypothesis regarding a computational limitation of distributed representations. On this view a network using distributed representations can unambiguously represent one item at a time, but superimposing two or more patterns over the same units can result in a blend pattern that is ambiguous in that there is no way to reconstruct the patterns from the blend. Localist representations do not suffer from this constraint.

localist codes have been found in deep recurrent networks trained to coactivate multiple items at the same time [27] as well as deep networks trained on items one at a time [28], highlighting that local codes are learned under a variety of network architectures and training conditions. Figure 1C provides an example of using single-unit recordings to reveal localist coding in a deep network.

In addition to the single-unit recording methods, selective units have been found in deep networks using a process called activation maximization. In this method, instead of presenting a set of meaningful images to a network and recording how individual units respond (as in the scatter plot method), the experimenter generates (through various algorithms) images that best activate specific target units. At the start of the process a random input pattern (noise) is presented to a network that only weakly activates the target unit, and, through an iterative process, images are synthesized that more strongly activate the unit. At the end of the process an image is generated that drives that unit more strongly than any other sampled image, and the question arising is what type of image is generated. If an interpretable image is synthesized it suggests that the unit has learned to code for meaningful high-level visual information, consistent with localist coding.

In fact, many reports of interpretable images have been documented [17], most often based on recordings from output units, but also from recordings of hidden units, as illustrated in Figure 1D. In a few cases the selective units found in deep networks have been called grandmother cells [28], but for the most part researchers do not make any psychological or neuroscientific claims. Instead, the authors are trying to understand how these networks work with the hope that this knowledge will inspire the creation of future models with better performance on applied tasks. Nevertheless, a better understanding of these conditions may also provide some insight into why some neurons selectively respond to meaningful inputs in hippocampus and cortex, and a growing number of single-unit recording studies have been carried out in deep networks with the goal of addressing psychological and neuroscience questions [29–31].

Symbolic Versus Non-Symbolic Computations

A key distinction between network types is whether or not they implement symbols. On one view, sometimes called ‘implementational connectionism’ [32] or ‘symbolic connectionism’ [33], networks are endowed with special mechanisms to support symbolic computations. By contrast, on the PDP approach neural networks may appear to support the computational capacities of symbolic systems under some conditions, but the underlying algorithms that mediate performance are simpler and non-symbolic [2,33].

To implement a symbolic neural network, words, objects, concepts, etc. need to be represented in long-term memory in a format that supports ‘compositionality’ such that complex

Figure 1. Different Methods of Displaying Selectivity. Different methods for depicting the selectivity of single units across a range of networks trained on a range of tasks. (A) An example of a ‘scatter plot’ developed by Berkeley *et al.* [14]. In this approach a separate scatterplot is created for each hidden unit, and each point in a scatterplot corresponds to the activation of a unit in response to a specific input. The level of unit activation is coded along the x axis, and distinct values are assigned to each point along the y axis. Berkeley *et al.* observed banding patterns after training, with inputs within a band sharing a meaningful feature. The scatter plot above depicts a single hidden unit in a model trained to categorize a set of logical problems as valid or invalid, and the points in highly active bands were all associated with the input feature ‘OR’. That is, this hidden unit is an ‘OR’ detector. (B) Scatter plot from [8] in which the points were labeled. This hidden unit responded selectively to words that contain the letter ‘i’. (C) Activation of a single hidden unit in a deep recurrent network trained to generate text after being trained with Leo Tolstoy’s *War and Peace* and the source code of the Linux Kernel [27]. The unit was highly active (indicated in lighter grey) after it generated an opening quote character, and remained active until the closing quote was output, at which point it turned off (indicated in darker grey). (D) Activation maximization method of depicting selectivity of single units in a deep convolutional network [12]. On this method images are synthesized that maximally activate a specific unit. The five images are the product of five different simulations of synthesizing images that maximally activate a single hidden unit. The fact that most of the images are interpretable (as a lighthouse in this case) suggests that the unit was tuned to code for a specific meaningful thing.

representations are composed from simpler parts in a regular fashion [9]. A key requirement of compositionality is that the parts maintain their identity in different contexts. This is achieved through a process of dynamically assigning the parts a role (or equivalently, assigning values to variables) to construct more complex representations, as described in [Box 1](#).

The main motivation for symbolic theories is the claim that context-independent representations are necessary for human-like generalization that occurs 'outside the training space' [11]. That is, symbolic systems can support generalization for new inputs that contain features that have not been trained on a given task. Marcus [11] gives the example of the identity function, $f(x) = x$, where it is necessary to assign a value (e.g., a number, word, object, etc.) to the variable x . If a person learns the identity function from a few examples, he or she can respond appropriately to an infinity of different inputs, including inputs that are highly dissimilar to the trained examples. For instance, after learning to respond 'one' to the spoken word 'one', and 'two' to the spoken word 'two', we have no difficulty generalizing to untrained numbers, or even untrained non-numbers presented in a different modality: given a picture of a duck, we can respond 'duck', or draw a duck. Crucially, the spoken words 'one' and 'two' share no input features with a picture of a duck, and nevertheless generalization is trivial, reflecting the human ability to generalize outside the training space. Generalization outside the training space is required for many high-level cognitive tasks [9,10] but, as detailed below, is also required for some memory and perceptual tasks. Generalization outside the training space is analogous to extrapolation, where predictions are made beyond the original observation range.

A Role for Symbolic Processes in PDP Networks

Despite the claim that symbols are needed for human-like generalization, non-symbolic models continue to be the overwhelming approach to studying human cognition in psychology and neuroscience. One reason for this is that PDP models often appear good at generalizing because they are typically tested 'within their training space'. That is, models are only tested on novel items that share all the relevant input features with trained items. For example, Botvinick and Plaut [21] emphasized that their recurrent PDP model of STM successfully recalled sequences of six letters despite the fact that the sequences were almost always novel (>99.9% of the time; simulation 1). However, although the specific sequences of six letters were almost always novel, the model had been trained many times with all letters in all list positions (e.g., although the sequence A-K-E-B-F-S may have been novel, the model was trained on many lists that contained A in the first position, K in the second position, etc.). When specific letters were excluded from specific list positions during training (e.g., the letter A was trained in all positions other than position 1), the model did poorly when tested on sequences that included these letters in these positions [34,35]. That is, the model failed to generalize outside the training space, exactly as critics of non-symbolic models would predict ([36] for a related finding).

Similarly, O'Reilly [37] developed a PDP model that identified horizontal and vertical bars presented in various positions and orientations on an input layer. The model was able to generalize to many unseen patterns based on limited training, and this was taken as a response to critics of non-symbolic connectionist theories. The authors wrote: 'Such results should help to counter the persistent claims that neural networks are incapable of producing systematic behavior based on statistical learning of the environment' (pp. 1230–1231). However, as noted by Dumas and Hummel [38], the training set in the simulations was 'judiciously chosen to span the space of all possible input and output vectors' (p. 79), and as such the model was not tested

outside the training space. Thus, again, the success of the model does not address the persistent criticisms raised by critics of non-symbolic models.

Perhaps the best-known example of PDP models generalizing comes from models of single-word reading that not only successfully name trained words but also many novel words (e.g., [39–41]; but see [42]). This was taken to undermine the dual-route model of naming that includes a symbolic grapheme–phoneme conversion system for the sake of generalizing to novel words (as well as localist codes for the sake of naming irregular words) [43]. However, once again, the PDP models were only tested within their training space (the models were only tested on novel words that included letters in trained positions).

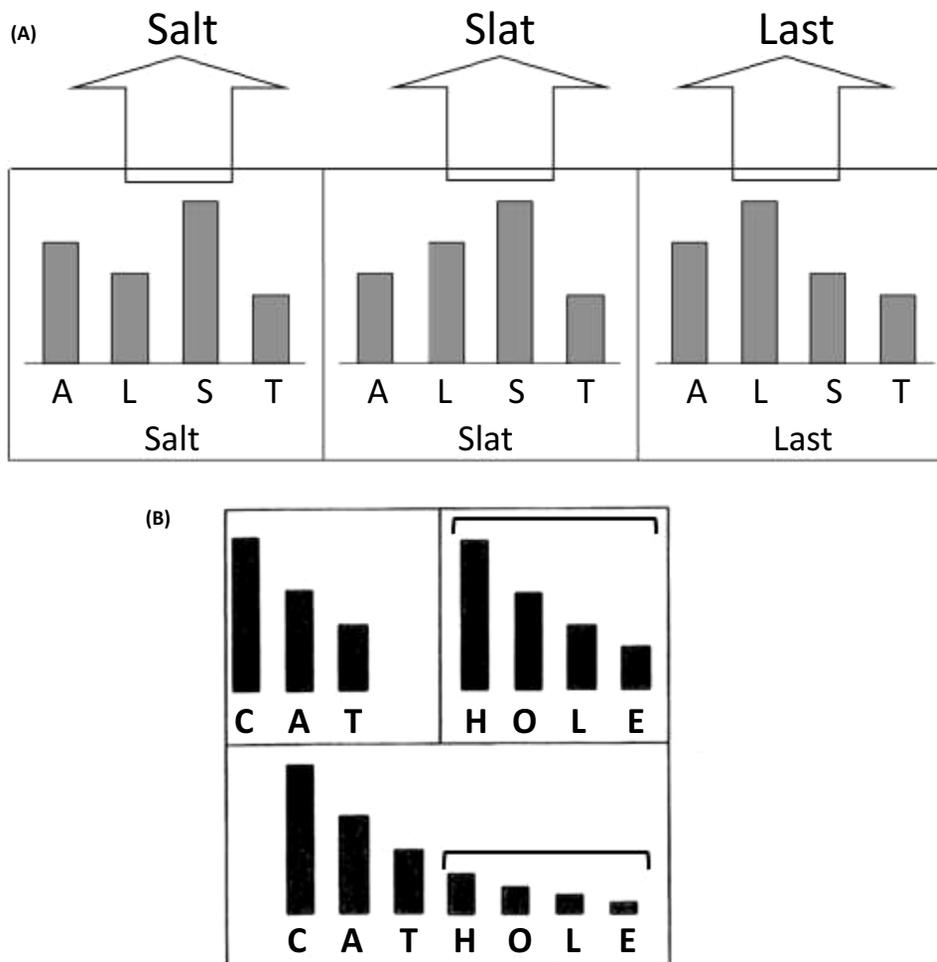
The limitation of non-symbolic models of word processing was highlighted in a recurrent PDP model that was trained to learn the orthographic forms of words [44]. The model was claimed to solve the challenge of identifying familiar words (e.g., COW) in novel contexts (e.g., PINKCOW), what Davis [45] called the ‘alignment problem’. To illustrate the problem, a PDP model that learns to identify the word COW on the basis of position-specific letter codes (e.g., C-in-position-1, O-in-position-2, and W-in-position-3) cannot identify the COW in BROWNCOW on the basis of the untrained C-in-position-4, O-in-position-5, and W-in-position-6 letter units (a case of testing outside the training space of the model). Thus, the claim of Sibley *et al.* that their recurrent PDP model solved the alignment problem and could ‘capture the commonalities between wordforms like CAT, SCAT, CATS, and SCATS’ (p. 742) was notable. However, the model was never actually tested on the alignment problem, and it was later shown that the model had no capacity to solve it [46]. Indeed, in a response article, Sibley *et al.* [47] wrote ‘We concede that according to their definition, the sequence encoder does not solve the alignment problem’ (p. 1189). Thus, again, a non-symbolic model failed to generalize outside the training space. To solve the alignment problem Davis [45,48] developed symbolic models of visual word identification that includes context-independent letter codes to identify familiar words in novel contexts. Figure 2 provides an illustration of how context-independent letter codes can be used to solve this problem.

A Role for Symbolic Processes in Deep Networks

Of course, the limited generalization capacities of small-scale PDP models does not guarantee that more powerful non-symbolic models will also fail. Thus, it is important to emphasize that more powerful deep networks show the same limitations.

For instance, Graves *et al.* [13] trained networks to perform a set of calculations on one graph (e.g., learning the shortest distance between two nodes) and then assessed whether the model could apply this knowledge to another graph (a graph of the London underground). The authors found that standard non-symbolic deep networks failed to generalize, and that it was necessary to add an external memory that implements a symbolic system to support generalization. This is how they motivate their hybrid network/symbolic system:

Modern computers separate computation and memory. Computation is performed by a processor, which can use an addressable memory to bring operands in and out of play. This confers two important benefits: the use of extensible storage to write new information and the ability to treat the contents of memory as variables. Variables are critical to algorithm generality: to perform the same procedure on one datum or another, an algorithm merely has to change the address it reads from. In contrast to computers, the computational and memory resources of artificial neural networks are mixed together in the network weights and neuron activity. This is a major liability: as the memory demands of a task increase, these



Trends in Cognitive Sciences

Figure 2. Coding Letters in a Context-Independent manner. (A) The same set of letters S, A, L, T are involved in coding for the words SALT, SLAT, and LAST, and the order of the letters is coded through the level of letter activation, with earlier letters being more active. (B) The use of context-independent codes provides a potential solution to the alignment problem. The pattern of letter of activation associated with the inputs CAT, HOLE, and CATHOLE is displayed. The crucial point to note is that the inputs HOLE and CATHOLE not only activate the same H, O, L, E letter codes but, in addition, the pattern of activation over these units is the same, with H being most active, followed by reduced activation of O, L, and E codes. Accordingly, if the model has learned to identify the word HOLE in isolation, then the model will be able to identify HOLE when presented in the novel context CATHOLE, solving the alignment problem [45]. This is not possible in parallel distributed processing (PDP) models in which letters are coded by position (e.g., there is H-in-position-1 in CATHOLE). This coding scheme, sometimes called a 'primacy gradient', was first developed by Grossberg [67] in the context of short-term memory (STM), and was similarly used by Page and Norris [68] in their symbolic model of STM.

networks cannot allocate new storage dynamically, nor easily learn algorithms that act independently of the values realized by the task variables . . . ' (p. 471). This passage could have been written by Jerry Fodor 30 years ago.

In fact, there are a growing number of 'memory networks' [49] that implement symbolic computations with specially designed memory systems that store items in a context-independent manner. The reason why the generalization limitations of non-symbolic models are becoming more widely appreciated is that computer scientists are trying to solve real-world problems that require more robust forms of generalization. The repeated successes of PDP

networks within the training space cannot mask the limitations of this approach any longer. Interestingly, a number of theorists traditionally associated with the non-symbolic PDP approach have recently been exploring ways to implement symbols in neural networks to improve generalization as well (e.g., [38,50]).

Furthermore, the reliance on non-symbolic representations is part of the reason why PDP and deep networks need to be trained so extensively. For example, deep networks of object identification, including deep **convolutional networks**, do not support robust translation invariance. That is, learning to identify an object in one retinal location does not immediately allow the model to generalize to distal retinal locations. As a consequence, it is necessary to train each trained object at many different locations [51,52], or add special modules that spatially manipulate the input patterns [53]. Apart from increasing the amount of training, this is not how human vision works [54]. A nice example of the advantages of learning using context-independent representations was recently reported by Lake *et al.* [55] who showed how models with symbolic capacities can support one-shot learning of letters, whereas deep networks need many training trials. Of course, in many cases it is not feasible to adequately sample the test space during training, and in these cases symbolic neural networks may be required.

How can Symbolic Processes Be Implemented in Neural Networks?

No doubt one of the reasons why there are still so few examples of symbolic theories in psychology and neuroscience is that it is much easier to build non-symbolic networks. Indeed, it is not immediately obvious how to implement symbolic processes in an artificial neural network, let alone in neural tissue.

As Gallistel and Balsam [56] write: 'Perhaps the biggest obstacle to neurobiologists' acceptance of the view that the brain stores information in symbolic form, just as does a computer, is our inability to imagine what this story might look like at the cellular and molecular level' (p. 142).

In fact, there have been different proposals over the years regarding how to implement symbolic computations in neural networks, and all proposals entail fundamental challenges to the PDP approach to theorizing (above and beyond implementing symbols). On one approach, all learning and computation takes place in the connection weights between units, and specialized modules and circuits are introduced to networks to encode and operate on context-independent representations [57]. This is a departure from the PDP approach according to which human cognition emerges from general learning algorithms operating on systems with minimal innate structure – the so-called 'emergentist view' [58].

The more radical approach to implementing symbolic computation is to reject the core PDP claim that all learning and computation takes place at the level of the connections between units [59]. For example, Gallistel and colleagues argue that symbolic computations are mediated by memories stored at the level of molecules within neurons [56], Hummel and colleagues [10,60] argue that neural synchrony is used for variable binding, and Davis [48] argues that delay-lines that alter the conduction times of neurons can be used to support symbolic computations. In fact, there is good evidence that learning and computation do take place outside the synapse [56], and, as detailed in [Box 2](#), recent evidence that myelin plasticity provides a biologically plausible mechanism for implementing delay-lines that have been used to support symbolic computations. In all symbolic theories, networks are endowed with additional resources to build the compositional representations and variable binding needed for symbolic computation.

Box 2. Learning and Computation Outside the Synapse with Myelin Plasticity

A fundamental claim of PDP theories is that all learning and computation is mediated by the connection weights between units [59]. This also characterizes almost all deep networks. However, recent studies have demonstrated another locus of learning and computation, namely the adaptive modification of myelin along axons that alters the neural conduction times of neurons – so-called myelin plasticity. For example, learning is associated with changes in myelin in brain regions relevant for performing a task [62], adaptive changes in myelin are used to insure the coincident arrival times of spikes on postsynaptic cells [63], and blocking myelin production impairs new learning in some tasks [64]. A recent review of myelin plasticity is given in [65].

The potential significance of identifying a second locus for learning and computation is difficult to overstate. First, the assumption that all learning and computation occurs at synapse has motivated artificial networks that compute with idealized units that are identical apart from the connections they make with other units. Consequently, models include units that take a fixed amount of time to pass information from input to output, and this in turn ensures that the relative timing of inputs on a postsynaptic unit are irrelevant. By contrast, actual neurons vary dramatically in their morphology, such that conduction times of communicating information between neurons vary dramatically. This is functionally relevant because the timing with which a postsynaptic neuron receives inputs (spikes) from multiple sources not only has a profound impact on the activation of the postsynaptic neurons as a result of temporal summation but the timing also impacts on learning due to spike-time-dependent learning [66]. Myelin plasticity provides a possible mechanism to adaptively modify the timing of neural signals so as to maximize the activation of postsynaptic neurons.

Second, myelin plasticity provides a possible implementation of delay-lines that have been proposed to support symbolic computations. Specifically, the symbolic spatial coding model of visual word identification [48] uses the connection weights between units to code for the identity of letters within words (in a context-independent manner) and uses delay-lines to dynamically code for the order of letters within words. Indeed, this model predicted a learning mechanism that adaptively alters the time it takes neurons to communicate information via delay-lines, precisely what myelin plasticity achieves. This model accounts for a large set of experimental results on visual word identification, and, crucially, solves the alignment problem.

Concluding Remarks and Future Perspectives

PDP models developed in the 1980s continue to have a profound impact on theories of mind and brain, and indeed rejection of localist codes and symbolic computations is the predominant view in psychology and neuroscience today. Given that PDP models are the precursor of deep networks, it is somewhat ironic that research with deep networks is providing some of the strongest arguments to date that localist representations and symbolic computations play an essential role in human cognition.

In the future, it will be important to explore in more detail the conditions in which artificial neural networks learn localist and distributed coding schemes, and see whether these findings relate to how the brain codes for information. In addition, the computational limitations of PDP and non-symbolic deep networks should motivate researchers to explore how symbolic representations and computations might be implemented in cognitive models and neural tissue. This may not only provide new insights into how the brain implements cognition but may also lead to more powerful artificial neural networks for solving more difficult artificial intelligence problems (see Outstanding Questions).

Acknowledgments

I would like to thank Ella Gale for comments on an earlier draft of manuscript. This work was supported by a grant from The Leverhulme Trust (RPG-2016-113) awarded to Jeffrey Bowers and Colin Davis. This project also received funding from the European Research Council (ERC) under the European Commission Horizon 2020 Research and Innovation Programme (grant agreement 741134).

References

1. Rumelhart, D.E. et al. (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (Vol. 1) Foundations, MIT Press
2. McClelland, J.L. et al. (1986) *Parallel Distributed Processing: Psychological and Biological Models* (Vol. 2), MIT Press

Outstanding Questions

What are the conditions in which PDP and deep networks learn highly selective codes? Will a better understanding of these conditions provide any insight into why some neurons respond in a highly selective manner? Or are artificial networks too different to make meaningful comparisons?

Are the highly selective neurons found in hippocampus and cortex consistent with localist models in psychology?

Are symbolic computations necessary to explain human cognition, or will more powerful non-symbolic networks suffice?

Does learning and computation occur outside of the links between units (synapses between neurons)?

Does the brain rely on different mechanisms to support symbolic computations in different domains? For instance, does the brain rely on synchrony of firing in some domains and delay-lines in others?

3. Graves, A. and Jaitly, N. (2014) Towards end-to-end speech recognition with recurrent neural networks. *Proceedings of Machine Learning Research* 32, 1764–1772
4. Hannun, A. et al. (2014) Deep speech: scaling up end-to-end speech recognition. *arXiv* <https://arxiv.org/abs/1412.5567>
5. He, K. et al. (2016) Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, IEEE
6. Krizhevsky, A. et al. (2012) ImageNet classification with deep convolutional neural networks. *Adv. Neural Inform. Process. Syst.* 25, 1097–1105
7. Huang, G. et al. (2016) Deep networks with stochastic depth. In *European Conference on Computer Vision*, pp. 646–661, IEEE
8. Bowers, J.S. et al. (2014) Neural networks learn highly selective representations in order to overcome the superposition catastrophe. *Psychol. Rev.* 121, 248–261
9. Fodor, J.A. and Pylyshyn, Z.W. (1988) Connectionism and cognitive architecture: a critical analysis. *Cognition* 28, 3–71
10. Hummel, J.E. and Holyoak, K.J. (1997) Distributed representations of structure: a theory of analogical access and mapping. *Psychol. Rev.* 104, 427–466
11. Marcus, G.F. (1998) Rethinking eliminative connectionism. *Cogn. Psychol.* 37, 243–282
12. Nguyen, A. et al. (2016) Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *Adv. Neural Inform. Process. Syst.* 29, 3387–3395
13. Graves, A. et al. (2016) Hybrid computing using a neural network with dynamic external memory. *Nature* 538, 471–476
14. Berkeley, I.S.N. et al. (1995) Density plots of hidden unit activations reveal interpretable bands. *Conn. Sci.* 7, 167–187
15. Bowers, J.S. et al. (2016) Why do some neurons in cortex respond to information in a selective manner? Insights from artificial neural networks. *Cognition* 148, 47–63
16. Rolls, E.T. (2017) Cortical coding. *Lang. Cogn. Neurosci.* 32, 316–329
17. Bowers, J.S. (2017) Grandmother cells and localist representations: a review of current thinking. *Lang. Cogn. Neurosci.* 32, 257–273
18. Bowers, J.S. (2009) On the biological plausibility of grandmother cells: implications for neural network theories in psychology and neuroscience. *Psychol. Rev.* 116, 220–251
19. Gubian, M. et al. (2017) Comparing single-unit recordings taken from a localist model to single-cell recording data: a good match. *Lang. Cogn. Neurosci.* 32, 380–391
20. Page, M. (2017) Localist models are compatible with information measures, sparseness indices, and complementary-learning systems in the brain. *Lang. Cogn. Neurosci.* 32, 366–379
21. Botvinick, M.M. and Plaut, D.C. (2006) Short-term memory for serial order: A recurrent neural network model. *Psychol. Rev.* 113, 201
22. Von der Malsburg, C. (1986) Am I thinking assemblies? In *Brain Theory* (Palm, G. and Aertsen, A., eds), pp. 161–176, Springer
23. Marr, D. et al. (1971) Simple memory: a theory for archicortex. *Philos. Trans. R. Soc. B Biol. Sci.* 262, 23–81
24. Grossberg, S. (1980) How does a brain build a cognitive code? *Psychol. Rev.* 87, 1–51
25. McCloskey, M. and Cohen, N.J. (1989) Catastrophic interference in connectionist networks: The sequential learning problem. *Psychol. Learn. Motiv.* 24, 109–165
26. Plaut, D.C. and McClelland, J.L. (2000) Stipulating versus discovering representations. *Behav. Brain Sci.* 23, 489–491
27. Karpathy, A. et al. (2015) Visualizing and understanding recurrent networks. *arXiv* <https://arxiv.org/abs/1506.02078>
28. Le, Q.V. (2012) Building high-level features using large scale unsupervised learning. In *Proceedings of the 29th International Conference on International Conference on Machine Learning*, pp. 507–514, Omnipress
29. Hannagan, T. et al. (2014) Deep learning of orthographic representations in baboons. *PLoS One* 9, e84843
30. Kheradpisheh, S.R. et al. (2016) STDP-based spiking deep neural networks for object recognition. *arXiv* <https://arxiv.org/abs/1611.01421>
31. Pinker, S. and Prince, A. (1988) On language and connectionism: analysis of a parallel distributed processing model of language acquisition. *Cognition* 28, 73–193
32. Holyoak, K.J. and Hummel, J.E. (2000) The proper treatment of symbols in a connectionist architecture. In *Cognitive Dynamics: Conceptual Change in Humans and Machines* (Dietrich, E. and Markman, A.B., eds), pp. 229–263, Erlbaum
33. McClelland, J.L. (2010) Emergence in cognitive science. *Top. Cogn. Sci.* 2, 751–770
34. Bowers, J.S. et al. (2009) A fundamental limitation of the conjunctive codes learned in PDP models of cognition: comment on Botvinick and Plaut (2006). *Psychol. Rev.* 116, 986–997
35. Bowers, J.S. et al. (2009) Postscript: more problems with Botvinick and Plaut's (2006) PDP model of short-term memory. *Psychol. Rev.* 116, 995–997
36. Kriete, T. et al. (2013) Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *Proc. Natl. Acad. Sci. U. S. A.* 110, 16390–16395
37. O'Reilly, R.C. (2001) Generalization in interactive networks: the benefits of inhibitory competition and Hebbian learning. *Neural Comput.* 13, 1199–1241
38. Dumas, L.A.A. and Hummel, J.E. (2005) Approaches to modeling human mental representations: what works, what doesn't and why. In *The Cambridge Handbook of Thinking and Reasoning* (Holyoak, K.J. and Morrison, R.G., eds), pp. 73–91, Cambridge University Press
39. Seidenberg, M.S. and McClelland, J.L. (1989) A distributed: developmental model of word recognition and naming. *Psychol. Rev.* 96, 523–568
40. Plaut, D.C. et al. (1996) Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychol. Rev.* 103, 56–115
41. Sibley, D.E. et al. (2010) Learning orthographic and phonological representations in models of monosyllabic and bisyllabic naming. *Eur. J. Cogn. Psychol.* 22, 650–668
42. Mousikou, P. et al. (2017) Moving beyond the monosyllable in models of skilled reading: Mega-study of disyllabic nonword reading. *J. Mem. Lang.* 93, 169–192
43. Coltheart, M. et al. (2001) DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychol. Rev.* 108, 204–256
44. Sibley, D. et al. (2008) Large-scale modeling of wordform learning and representation. *Cogn. Sci. A Multidiscip. J.* 32, 741–754
45. Davis, C.J. (2001) The self-organising lexical acquisition and recognition (SOLAR) model of visual word recognition. *Dissertation Abstracts International: Section B Sciences and Engineering* 62, 594
46. Bowers, J.S. and Davis, C.J. (2009) Learning representations of wordforms with recurrent networks: comment on Sibley, Kello, Plaut, & Elman (2008). *Cogn. Sci.* 33, 1183–1186
47. Sibley, D.E. et al. (2009) Sequence encoders enable large-scale lexical modeling: reply to Bowers and Davis (2009). *Cogn. Sci.* 33, 1187–1191
48. Davis, C.J. (2010) The spatial coding model of visual word identification. *Psychol. Rev.* 117, 713–758
49. Weston, J. et al. (2014) Memory networks. *arXiv* <https://arxiv.org/abs/1410.3916>
50. Botvinick, M.M. and Cohen, J.D. (2014) The computational and neural basis of cognitive control: Charted territory and new frontiers. *Cogn. Sci.* 38, 1249–1285
51. Dandurand, F. et al. (2013) Computational models of location-invariant orthographic processing. *Conn. Sci.* 25, 1–26
52. Di Bono, M.G. and Zorzi, M. (2013) Deep generative learning of location-invariant visual word recognition. *Front. Psychol.* 4, 635

53. Jaderberg, M. *et al.* (2015) Spatial transformer networks. *Adv. Neural Inform. Process. Syst.* 28, 2017–2025
54. Bowers, J.S. *et al.* (2016) The visual system supports online translation invariance for object identification. *Psychon. Bull. Rev.* 23, 432–438
55. Lake, B.M. *et al.* (2015) Human-level concept learning through probabilistic program induction. *Science* 350, 1332–1338
56. Gallistel, C.R. and Balsam, P.D. (2014) Neurobiology of learning and memory, time to rethink the neural mechanisms of learning and memory. *Neurobiol. Learn. Mem.* 108, 136–144
57. Kriete, T. *et al.* (2013) Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *Proc. Natl. Acad. Sci. U. S. A.* 110, 16390–16395
58. McClelland, J.L. *et al.* (2010) Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends Cogn. Sci.* 14, 348–356
59. Rogers, T.T. and McClelland, J.L. (2014) Parallel distributed processing at 25: further explorations in the microstructure of cognition. *Cogn. Sci.* 38, 1024–1077
60. Hummel, J.E. and Biederman, I. (1992) Dynamic binding in a neural network for shape recognition. *Psychol. Rev.* 99, 480–517
61. McClelland, J.L. and Rumelhart, D.E. (1981) An interactive activation model of context effects in letter perception. I. An account of basic findings. *Psychol. Rev.* 88, 375–407
62. Bengtsson, S.L. *et al.* (2005) Extensive piano practicing has regionally specific effects on white matter development. *Nat. Neurosci.* 8, 1148–1150
63. Seidl, A.H. *et al.* (2010) Mechanisms for adjusting interaural time differences to achieve binaural coincidence detection. *J. Neurosci.* 30, 70–80
64. McKenzie, I.A. *et al.* (2014) Motor skill learning requires active central myelination. *Science* 346, 318–322
65. Purger, D. *et al.* (2016) Myelin plasticity in the central nervous system. *Neuropharmacology* 110, 563–573
66. Dan, Y. and Poo, M.M. (2004) Spike timing-dependent plasticity of neural circuits. *Neuron* 44, 23–30
67. Grossberg, S. (1978) Behavioral contrast in short term memory: serial binary memory models or parallel continuous memory models? *J. Math. Psychol.* 17, 199–219
68. Page, M. and Norris, D. (1998) The primacy model: a new model of immediate serial recall. *Psychol. Rev.* 105, 761–781