

## 2 Connectionist Models of Cognition

Michael S. C. Thomas and James L. McClelland

### 2.1 Introduction

In this chapter, computer models of cognition that have focused on the use of neural networks are reviewed. These architectures were inspired by research into how computation works in the brain, and particularly the observation that large, densely connected networks of relatively simple processing elements can solve some complex tasks fairly easily in a modest number of sequential steps. Subsequent work has produced models of cognition with a distinctive flavor. Processing is characterized by patterns of activation across simple processing units connected together into complex networks. Knowledge is stored in the strength of the connections between units. It is for this reason that this approach to understanding cognition has gained the name of *connectionism*.

Since the first edition of this volume, it has become apparent that the field has entered the third age of artificial neural network research. The first began in the 1930s and 1940s, part of the genesis of the first formal theories of computation; the second arose in the 1980s and 1990s with Parallel Distributed Processing models of cognition; and the third emerged in the mid-2000s with advances in “deep” neural networks. Transition between the ages has been triggered by new insights into how to create and train more powerful artificial neural networks.

### 2.2 Background

Over the last forty years, connectionist modeling has formed an influential approach to the computational study of cognition. It is distinguished by its appeal to principles of neural computation to inspire the primitives that are included in its cognitive level models. Also known as artificial neural network (ANN) or parallel distributed processing (PDP) models, connectionism has been applied to a diverse range of cognitive abilities, including models of memory, attention, perception, action, language, concept formation, and reasoning (see, e.g., Houghton, 2005; Joanisse & McClelland, 2015; Mayor, Gomez, Chang, & Lupyan, 2014). While many of these models seek to capture adult function, connectionism places an emphasis on learning internal representations. This has led to an increasing focus on developmental phenomena

and the origins of knowledge. Although, at its heart, connectionism comprises a set of computational formalisms, it has spurred vigorous theoretical debate regarding the nature of cognition. Some theorists have reacted by dismissing connectionism as mere implementation of preexisting verbal theories of cognition, while others have viewed it as a candidate to replace the Classical Computational Theory of Mind and as carrying profound implications for the way human knowledge is acquired and represented; still others have viewed connectionism as a sub-class of statistical models involved in universal function approximation and data clustering.

The chapter begins by placing connectionism in its historical context, leading up to its formalization in Rumelhart and McClelland's two-volume *Parallel Distributed Processing* (1986) written in combination with members of the Parallel Distributed Processing Research Group. The innovations that then triggered the emergence of deep networks are indicated. Next, there is a discussion of three important foundational cognitive models that illustrate some of the key properties of connectionist systems and indicate how the novel theoretical contributions of these models arose from their key computational properties. These three models are the Interactive Activation model of letter recognition (McClelland & Rumelhart, 1981; Rumelhart and McClelland, 1982), Rumelhart and McClelland's model of the acquisition of the English past tense (1986), and Elman's simple recurrent network for finding structure in time (1991). The chapter finishes by considering how connectionist modeling has influenced wider theories of cognition, and how in the future, connectionist modeling of cognition may progress by integrating further constraints from neuroscience and neuroanatomy.

### 2.2.1 Historical Context

Connectionist models draw inspiration from the notion that the information processing properties of neural systems should influence theories of cognition. The possible role of neurons in generating the mind was first considered not long after the existence of the nerve cell was accepted in the latter half of the nineteenth century (Cobb, 2020). Early neural network theorizing can therefore be found in some of the associationist theories of mental processes prevalent at the time (e.g., Freud, 1895; James, 1890; Meynert, 1884; Spencer, 1872). However, this line of theorizing was quelled when Lashley presented data appearing to show that the performance of the brain degraded gracefully depending only on the quantity of damage. This argued against the specific involvement of neurons in particular cognitive processes (see, e.g., Lashley, 1929).

In the 1930s and 1940s, there was a resurgence of interest in using mathematical techniques to characterize the behavior of networks of nerve cells (e.g., Rashevsky, 1935). This culminated in the work of McCulloch and Pitts (1943) who characterized the function of simple networks of binary threshold neurons in terms of logical operations. In his 1949 book *The Organization of Behavior*,

Donald Hebb proposed a cell assembly theory of cognition, including the idea that specific synaptic changes might underlie psychological principles of learning. A decade later, Rosenblatt (1958, 1962) formulated a learning rule for two-layered neural networks, demonstrating mathematically that the *perceptron convergence rule* could adjust the weights connecting an input layer and an output layer of simple neurons to allow the network to associate arbitrary binary patterns (see also Novikoff, 1962). With this rule, learning converged on the set of connection values necessary to acquire any two-layer-computable function relating a set of input–output patterns. Unfortunately, Minsky and Papert (1969) demonstrated that the set of two-layer computable functions was somewhat limited – that is, these simple artificial neural networks were not particularly powerful devices. While more computationally powerful networks could be described, there was no algorithm to learn the connection weights of these systems. Such networks required the postulation of additional internal or “hidden” processing units, which could adopt intermediate representational states in the mapping between input and output patterns. An algorithm (backpropagation) able to learn these states was discovered independently several times. A key paper by Rumelhart, Hinton, and Williams (1986) demonstrated the usefulness of networks trained using backpropagation for addressing key computational and cognitive challenges facing neural networks.

In the 1970s, serial processing and the Von Neumann computer metaphor dominated cognitive psychology, relying heavily on symbolic representations (Newell, 1980). Nevertheless, a number of researchers continued to work on the computational properties of neural systems. Some of the key themes identified by these researchers include the role of competition in processing and learning (e.g., Grossberg, 1976a; Kohonen, 1984), and the use of hierarchically organized bi-directional connectivity for perceptual inference in adaptive competitive interactive systems (Grossberg, 1976b).

Researchers also began to explore the properties of distributed representations (e.g., Anderson, 1977; Hinton & Anderson, 1981), and the possibility of content addressable memory in networks with attractor states, formalized using the mathematics of statistical physics (Hopfield, 1982). A fuller characterization of the many historical influences in the development of connectionism can be found in Rumelhart, McClelland and the PDP Research Group (1986, chapter 1), Bechtel and Abrahamsen (1991), McLeod, Plunkett, and Rolls (1998), and O’Reilly and Munakata (2000).

Backpropagation networks prompted an explosion of models targeting simplified versions of problem domains from language and cognition. But it seemed for many years that such networks could not readily scale to complex, real-world problems such as natural language processing or vision. Once again, the issue was not that it was impossible to describe sufficiently powerful networks, but that such networks were not trainable using the available tools. This time, instead of a single breakthrough, this barrier was overcome by several convergent developments. These included several architectural and processing enhancements, the availability of much greater computational power, and the

availability of large data sets to train the models (LeCun, Bengio, & Hinton, 2015). Now, instead of shallow networks typically containing only three layers (input, hidden, and output), networks with tens or even hundreds of layers (hence, “deep”) could be trained to solve complex problems. The latest deep neural networks are now applied to problems such as visual object recognition, speech recognition, and natural language processing, sometimes showing near human or even super-human levels of performance (Kriegeskorte, 2015; Storrs & Kriegeskorte, 2019; see also Chapter 9 in this handbook).

Figure 2.1 depicts a selective schematic of this history and demonstrates the multiple types of neural network system that have latterly come to be used in building models of cognition. While diverse, they are unified on the one hand by the proposal that cognition comprises processes of constraint satisfaction, energy minimization and pattern recognition, and on the other that adaptive processes construct the microstructure of these systems, primarily by adjusting the strengths of connections among the neuron-like processing units involved in a computation.

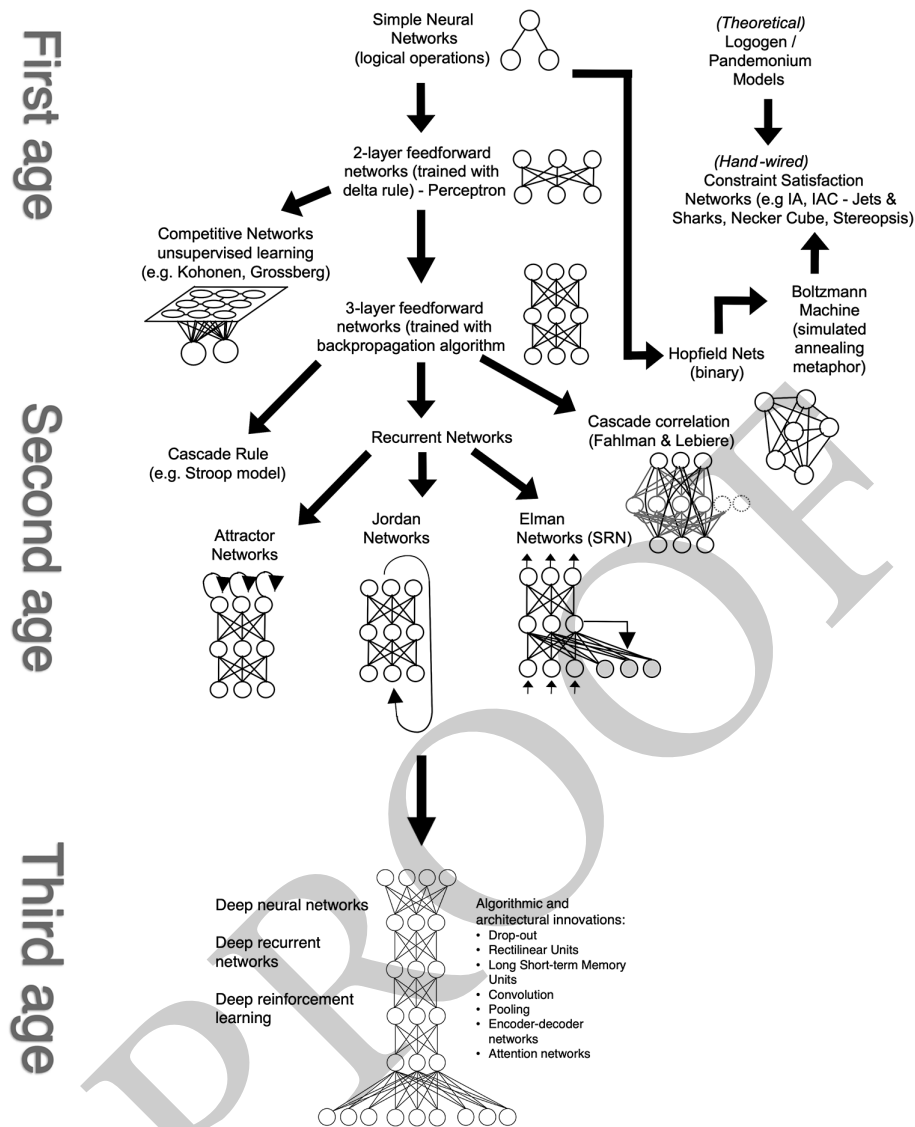
### 2.2.2 Key Properties of Connectionist Models

Connectionism starts with the following inspiration from neural systems: computations will be carried out by a set of simple processing units operating in parallel and affecting each other’s activation states via a network of weighted connections. Rumelhart, Hinton, and McClelland (1986) identified seven key features that would define a general framework for connectionist processing.

The first feature is the set of processing units  $u_i$ . In a cognitive model, these may be intended to represent individual concepts (such as letters or words), or they may simply be abstract elements over which meaningful patterns can be defined. Processing units are often distinguished into input, output, and hidden units. In associative networks, input and output units have states that are defined by the task being modeled (at least during training), while hidden units are free parameters whose states may be determined as necessary by the learning algorithm.

The second feature is a state of activation ( $a$ ) at a given time ( $t$ ). The state of a set of units is usually represented by a vector of real numbers  $a(t)$ . These may be binary or continuous numbers, bounded or unbounded. A frequent assumption is that the activation level of simple processing units will vary continuously between the values 0 and 1.

The third feature is a pattern of connectivity. The strength of the connection between any two units will determine the extent to which the activation state of one unit can affect the activation state of another unit at a subsequent time point. The strength of the connections between unit  $i$  and unit  $j$  can be represented by a matrix  $W$  of weight values  $w_{ij}$ . Multiple matrices may be specified for a given network if there are connections of different types. For example, one matrix may specify excitatory connections between units and a second may specify inhibitory connections. Potentially, the weight matrix allows every unit



**Figure 2.1** A simplified schematic showing the historical evolution of neural network architectures. Simple binary networks (McCulloch & Pitts, 1943) are followed by two-layer feedforward networks (perceptrons; Rosenblatt, 1958). Three subtypes then emerge: feedforward networks (Rumelhart & McClelland, 1986), competitive or self-organizing networks (e.g., Grossberg, 1976a; Kohonen, 1984), and symmetrically connected energy-minimization networks (Hinton & Sejnowski, 1986; Hopfield, 1982). Adaptive interactive networks have precursors in detector theories of perception (Logogen: Morton, 1969; Pandemonium: Selfridge, 1959) and hard-wired interactive models (Interactive Activation: McClelland & Rumelhart, 1981; Interactive Activation and Competition: McClelland, 1981; Stereopsis: Marr & Poggio, 1976; Necker cube: Feldman, 1981), and Grossberg provided an early adaptive learning rule for such systems (Grossberg, 1976b). Feedforward pattern

to be connected to every other unit in the network. Typically, units are arranged into layers (e.g., input, hidden, output) and layers of units are fully connected to each other. For example, in a three-layer feedforward architecture where activation passes in a single direction from input to output, the input layer would be fully connected to the hidden layer and the hidden layer would be fully connected to the output layer.

The fourth feature is a rule for propagating activation states throughout the network. This rule takes the vector  $a(t)$  of output values for the processing units sending activation and combines it with the connectivity matrix  $W$  to produce a summed or net input into each receiving unit. The net input to a receiving unit is produced by multiplying the vector and matrix together, so that

$$net_i = W \times a(t) = \sum_j w_{ij} a_j \quad (2.1)$$

The fifth feature is an activation rule to specify how the net inputs to a given unit are combined to produce its new activation state. The function  $F$  derives the new activation state

$$a_i(t+1) = F(net_i(t)) \quad (2.2)$$

For example,  $F$  might be a threshold so that the unit becomes active only if the net input exceeds a given value. Other possibilities include linear, Gaussian, and sigmoid functions, depending on the network type. Sigmoid is perhaps the most common, operating as a smoothed threshold function that is also differentiable. It is often important that the activation function be differentiable because learning seeks to improve a performance metric that is assessed via the activation state while learning itself can only operate on the connection weights. The effect of weight changes on the performance metric therefore depends to some extent on the activation function, and the learning algorithm encodes this fact by including the derivative of that function (see below).

The sixth key feature of connectionist models is the algorithm for modifying the patterns of connectivity as a function of experience. Virtually all learning rules for PDP models can be considered a variant of the Hebbian learning rule (Hebb, 1949). The essential idea is that a weight between two units should be

---

**Caption for Figure 2.1** (cont.) *associators have been extended to three or more layers with the introduction of backpropagation (Rumelhart, Hinton & Williams, 1986), and have produced multiple subtypes used in modeling dynamic aspects of cognition: these include cascaded feedforward networks (e.g., Cohen, Dunbar, & McClelland, 1990) and attractor networks in which states cycle into stable configurations (e.g., Plaut & McClelland, 1993); for processing sequential information, recurrent networks (Elman, 1991; Jordan, 1986); for systems that alter their structure as part of learning, constructivist networks (e.g., cascade correlation: Fahlman & Lebiere, 1990; Shultz, 2003). Since the early 2000s, deep neural networks have emerged, characterized by multiple layers of hidden units (LeCun, Bengio, & Hinton, 2015).*

altered in proportion to the units' correlated activity. For example, if a unit  $u_i$  receives input from another unit  $u_j$ , then if both are highly active, the weight  $w_{ij}$  from  $u_j$  to  $u_i$  should be strengthened. In its simplest version, the rule is

$$\Delta w_{ij} = \eta a_i a_j \quad (2.3)$$

where  $\eta$  is the constant of proportionality known as the learning rate. Where an external target activation  $t_i(t)$  is available for a unit  $i$  at time  $t$ , this algorithm is modified by replacing  $a_i$  with a term depicting the disparity of unit  $u_i$ 's current activation state  $a_i(t)$  from its desired activation state  $t_i(t)$  at time  $t$ , so forming the delta rule:

$$\Delta w_{ij} = \eta(t_i(t) - a_i(t))a_j \quad (2.4)$$

However, when hidden units are included in networks, no target activation is available for these internal parameters. The weights to such units may be modified by variants of the Hebbian learning algorithm (e.g., Contrastive Hebbian; Hinton, 1989; see Xie & Seung, 2003) or by the backpropagation of error signals from the output layer.

Backpropagation makes it possible to determine, for each connection weight in the network, what effect a change in its value would have on the overall network error. The policy for changing the strengths of connections is simply to adjust each weight in the direction (up or down) that would tend to reduce the error, by an amount proportional to the size of the effect the adjustment will have. If there are multiple layers of hidden units remote from the output layer, this process can be followed iteratively: first error derivatives are computed for the hidden layer nearest the output layer; from these, derivatives are computed for the next deepest layer into the network, and so forth. On this basis, the backpropagation algorithm serves to modify the pattern of weights in powerful multilayer networks. It alters the weights to each deeper layer of units in such a way as to reduce the error on the output units (see Rumelhart, Hinton, & Williams, 1986, for the derivation). The weight change algorithm can be formulated by analogy to the delta rule as shown in Equation 2.4. For each deeper layer in the network, the central term that represents the disparity between the actual and target activation of the units is modified. Assuming  $u_i$ ,  $u_h$ , and  $u_o$  are input, hidden, and output units in a three-layer feedforward network, the algorithm for changing the weight from hidden to output unit is:

$$\Delta w_{oh} = \eta(t_o - a_o)F'(net_o)a_h \quad (2.5)$$

where  $F'(net)$  is the derivative of the activation function of the units (e.g., for the sigmoid activation function,  $F'(net_o) = a_o(1 - a_o)$ ). The term  $(t_o - a_o)$  is proportional to the negative of the partial derivative of the network's overall error with respect to the activation of the output unit, where the error  $E$  is given by  $E = \sum_o (t_o - a_o)^2$ .

The derived error term for a unit at the hidden layer is based on the derivative of the hidden unit's activation function, times the sum across all the connections

from that hidden unit to the output later of the error term on each output unit weighted by the derivative of the output unit's activation function  $(t_o - a_o)F'(net_o)$  times the weight connecting the hidden unit to the output unit:

$$F'(net_h) \sum_o (t_o - a_o) F'(net_o) w_{oh} \quad (2.6)$$

The algorithm for changing the weights from the input to the hidden layer is therefore:

$$\Delta w_{hi} = \eta F'(net_h) \sum_o (t_o - a_o) F'(net_o) w_{oh} a_i \quad (2.7)$$

It is interesting that the above computation can be construed as a backward pass through the network, similar in spirit to the forward pass that computes activations in that it involves propagation of signals across weighted connections, this time from the output layer back toward the input. The backward pass, however, involves the propagation of error derivatives rather than activations.

It should be emphasized that a very wide range of variants and extensions of Hebbian and error-correcting algorithms have been introduced in the connectionist learning literature. Most importantly, several variants of backpropagation have been developed for training recurrent networks, that is, those in which activation can cycle around loops (Williams & Zipser, 1995); and several algorithms (including the Contrastive Hebbian Learning algorithm and O'Reilly's 1998 LEABRA algorithm) have addressed some of the concerns that have been raised regarding the biological plausibility of backpropagation construed in its most literal form (O'Reilly & Munakata, 2000).

One challenge of training deep neural networks, with many layers of hidden units, is called the vanishing gradient problem (Hochreiter, 1991). As has been seen, the change to each layer of weights extending deeper into the network (that is, further from the output, closer to the input) depends on the extent to which each weight contributes to the error at the output layer, scaled by the gradient of the activation function at each layer of units above. Since for many activation functions, such as the sigmoid, the gradient falls between 0 and 1, this results in the multiplication of several numbers each less than one: potentially it produces very small weight change at deeper layers, slowing down learning. A parallel problem exists for recurrent networks, where each pass through the recurrent loop involves multiplying the weight change by another activation function derivative (Hochreiter et al., 2001). Equivalently, weight changes can be very small in response to information separated by several recurrent passes through the network. Indeed, in practice, the vanishing gradient problem may be more serious for recurrent networks than feedforward networks, since the identical weights are involved in each iteration around a recurrent loop, guaranteeing exponential decay of the error signal. Together with other challenges (such as the disappearing signal problem, where many intermediate layers of initially randomized weights create noise that makes it hard to detect input-output relationships), the result was a limitation in the scalability of backpropagation networks



to the depth required to solve complex real-world problems, such as natural language processing or vision.

Several innovations subsequently made the training of deep neural networks viable, aided by large increases in computational power (perhaps a million-fold since the early 1990s; Schmidhuber, 2015). These included *drop out*, randomly disabling a subset of input units and hidden units on a given pattern presentation, which aids learning of more robust, generalizable input–output functions (Srivastava et al., 2014); *rectified linear units*, activation functions that are linear when their net input is greater than zero, but deactivated when less than zero – the larger, consistent gradient reduces the vanishing gradient problem deeper in the network (Hahnloser et al., 2000); and for image processing, *convolution networks*, which use structures analogous to visual receptive fields, serving to duplicate what is learned about useful visual features in one area of an input retina to other areas, so that location-invariant recognition is possible when this information is pooled (e.g., Krizhevsky, Sutskever, & Hinton, 2012).

For natural language processing, an important innovation was the use of *long short-term memory (LSTM) units* in recurrent networks. These units can hold information over as many recurrent cycles as necessary before feeding it into a computation, enabling the learning of dependencies further separated in time (Hochreiter & Schmidhuber, 1997). However, LSTMs only partially alleviated the central problem facing recurrent networks, which is that contextual information still had to be funneled through a very narrow bottleneck (a “context” vector of the same length as the previous hidden state in a simple recurrent network). The breakthroughs in natural language processing that attracted public notice in 2016 with the introduction of the Google Neural Machine Translation system depend on an innovation called Query Based Attention (see McClelland, Hill, Rudolph, Baldrige, & Schuetze, 2020, for an explanation of this mechanism; and also Chapter 9 in this handbook). Broadly, the attention mechanism stores multiple versions of the preceding context and then learns to differently weight them when predicting the output – in effect, helping to solve the problem of what in the input sequence goes with what in the output sequence.

Another important development has been the use of weak supervisory signals, in the form of reward or reinforcement signals, which only indicate whether a network is right or wrong, instead of specifying exactly what it should do. While such reinforcement-based approaches have been investigated within a neural network framework for decades (e.g., Sutton & Barto, 1981), their potential to address cognitively interesting problems stems from further innovations enabled by the massive scale of computation that has only been available recently. For instance, breakthroughs in playing games such as chess or Go stem from architectures enabled by increased computational power, which allows a system to play games with itself millions of times to identify the sequences of moves that produce the best possible outcomes. These innovations are further described in Chapter 10 in this handbook.

The seventh and last general feature of connectionist networks is a representation of the environment with respect to the system. This is assumed to consist

of a set of externally provided events or a function for generating such events. An event may be a single pattern, such as a visual input; an ensemble of related patterns, such as the spelling of a word and its corresponding sound and/or meaning; or a sequence of inputs, such as the words in a sentence. A range of policies have been used for specifying the order of presentation of the patterns, including sweeping through the full set to random sampling with replacement. The selection of patterns to present may vary over the course of training but is often fixed. Where a target output is linked to each input, this is usually assumed to be simultaneously available.

Two points are of note in the translation between PDP network and cognitive model. First, a representational scheme must be defined to map between the cognitive domain of interest and a set of vectors depicting the relevant informational states or mappings for that domain. Second, in many cases, connectionist models are addressed to aspects of higher-level cognition, where it is assumed that the information of relevance is more abstract than sensory or motor codes. This has meant that the models often leave out details of the transduction of sensory and motor signals, using input and output representations that are already somewhat abstract. The same principles at work in higher-level cognition are also held to be at work in perceptual and motor systems, and indeed there is also considerable connectionist work addressing issues of perception and action, though these will not be the focus of the present chapter.

### 2.2.3 Neural Plausibility

It is a historical fact that most connectionist modelers have drawn their inspiration from the computational properties of neural systems. However, it has become a point of controversy whether these “brain-like” systems are indeed neurally plausible. If they are not, should they instead be viewed as a class of statistical function approximators? And if so, should not the ability of these models to simulate patterns of human behavior be judged in the context of the large number of free parameters they contain (e.g., in the weight matrix) (Green, 1998)?

Neural plausibility should not be the primary focus for a consideration of connectionism. The advantage of connectionism, according to its proponents, is that it provides *better theories of cognition*. Nevertheless, this issue will be briefly dealt with since it pertains to the origins of connectionist cognitive theory. In this area, two sorts of criticism have been leveled at connectionist models. The first is to maintain that many connectionist models either include properties that are not neurally plausible and/or omit other properties that neural systems appear to have (e.g., Crick, 1989). Some connectionist researchers have responded to this first criticism by endeavoring to show how features of connectionist systems might in fact be realized in the neural machinery of the brain. For example, the backward propagation of error across the same connections



that carry activation signals is generally viewed as biologically implausible. However, a number of authors have shown that the difference between activations computed using standard feedforward connections and those computed using standard return connections can be used to derive the crucial error derivatives required by backpropagation (Hinton & McClelland, 1988; O'Reilly, 1996), even indeed if those return connections simply have random weights (Lillicrap et al., 2016). It is widely held that connections run bidirectionally in the brain, as required for this scheme to work. Under this view, backpropagation may be shorthand for a Hebbian-based algorithm that uses bidirectional connections to spread error signals throughout a network (Xie & Seung, 2003). This view was encapsulated in Lillicrap et al.'s (2020) proposal that the brain's feedback connections induce neural activities whose differences can be used to locally approximate error signals and drive effective learning in deep networks in the brain. Other researchers have argued that the apparent limited biological plausibility of backpropagation stems not from the algorithm per se but to the lack of temporal extension of processing in its usual implementation (specifically the instantaneous mapping from the input to output) (e.g., Betti & Gori, 2020; Scellier & Bengio, 2019).

Other connectionist researchers have responded to the first criticism by stressing the cognitive nature of current connectionist models. Most of the work in developmental neuroscience addresses behavior at levels no higher than cellular and local networks, whereas cognitive models must make contact with the human behavior studied in psychology. Some simplification is therefore warranted, with neural plausibility compromised under the working assumption that the simplified models share the same flavor of computation as actual neural systems. Connectionist models have succeeded in stimulating a great deal of progress in cognitive theory – and sometimes generating radically different proposals to the previously prevailing symbolic theory – just given the set of basic computational features outlined in the preceding section.

The second type of criticism leveled at connectionism questions why, as Davies (2005) put it, connectionist models should be reckoned any more plausible as putative descriptions of cognitive processes just because they are “brain-like.” Under this view, there is independence between levels of description because a given cognitive level theory might be implemented in multiple ways in different hardware. Therefore the details of the hardware (in this case, the brain) need not concern the cognitive theory. This functionalist approach, most clearly stated in Marr's three levels of description (computational, algorithmic, and implementational; see Marr, 1982) has been repeatedly challenged (see, e.g., Mareschal et al., 2007; Rumelhart & McClelland, 1985). The challenge to Marr goes as follows. While, according to computational theory, there may be a principled independence between a computer program (the “software”) and the particular substrate on which it is implemented (the “hardware”), in practical terms, different sorts of computation are easier or harder to implement on a given substrate. Since computations have to be delivered in real time as the

individual reacts with his or her environment, in the first instance cognitive-level theories should be constrained by the computational primitives that are most easily implemented on the available hardware; human cognition should be shaped by the processes that work best in the brain.

The relation of connectionist models to symbolic models has also proved controversial. A full consideration of this issue is beyond the scope of the current chapter. Suffice to say that because the connectionist approach now includes a diverse family of models, there is no single answer to this question. Smolensky (1988) argued that connectionist models exist at a lower (but still cognitive) level of description than symbolic cognitive theories, a level that he called the *sub-symbolic*. Connectionist models have sometimes been put forward as a way to implement symbolic production systems on neural architectures (e.g., Touretzky & Hinton, 1988). At other times, connectionist researchers have argued that their models represent a qualitatively different form of computation: while under certain circumstances, connectionist models might produce behavior approximating symbolic processes, it is held that human behavior often only approximates the characteristics of symbolic systems rather than directly implementing them. That is, when human behavior is (approximately) rule-following, it need not be rule-driven. Furthermore, connectionist systems incorporate additional properties characteristic of human cognition, such as content addressable memory, context-sensitive processing, and graceful degradation under damage or noise. Under this view, symbolic theories are approximate descriptions rather than actual characterizations of human cognition. Connectionist theories should replace them because they both capture subtle differences between human behavior and symbolic characterizations, and because they provide a specification of the underlying causal mechanisms (van Gelder, 1991).

This strong position has prompted criticisms that connectionist models are insufficiently powerful to account for certain aspects of human cognition – in particular those areas best characterized by symbolic, syntactically driven computations (Fodor & Pylyshyn, 1988; Lake et al, 2017; Marcus, 2001). Again, however, the characterization of human cognition in such terms is highly controversial; close scrutiny of relevant aspects of language – the ground on which the dispute has largely been focused – lends support to the view that the systematicity assumed by proponents of symbolic approaches is overstated, and that the actual characteristics of language are well matched to the characteristics of connectionist systems (Bybee & McClelland, 2005; Kollias & McClelland, 2013; McClelland, Plaut, Gotts, & Maia, 2003). Furthermore, recent breakthroughs in machine language processing now demonstrate that aspects of structure can emerge in powerful ways from neural networks that have been trained on large text corpora (see Section 2.3.3). Nevertheless, explanations of explicitly symbolic ways of thinking remain an  debate, including behaviors such as generalization over variables  that are less readily delivered by connectionist architectures.

## 2.2.4 The Relationship Between Connectionist Models and Bayesian Inference

Since the early 1980s, it has been apparent that there are strong links between the calculations carried out in connectionist models and key elements of Bayesian calculations (McClelland, 2013). It was noted, first of all, that units can be viewed as playing the role of probabilistic hypotheses; that weights and biases play the role of conditional probability relations between hypotheses and prior probabilities, respectively; and that if connection weights and biases have the correct values, the logistic activation function sets the activation of a unit to its posterior probability given the evidence represented on its inputs. A second and more important observation is that, in stochastic neural networks (Boltzmann Machines and Continuous Diffusion Networks; Hinton & Sejnowski, 1986; Movellan & McClelland, 1993) a network's state over all of its units can represent a constellation of hypotheses about an input; and (if the weights and the biases are set correctly) that the probability of finding the network in a particular state is monotonically related to the probability that the state is the correct interpretation of the input. The exact nature of the relation depends on a parameter called temperature; if set to one, the probability that the network will be found in a particular state exactly matches its posterior probability. When temperature is gradually reduced to zero, the network will end up in the most probable state, thus performing optimal perceptual inference (Hinton & Sejnowski, 1983). It is also known that back-propagation can learn weights that allow Bayes-optimal estimation of outputs given inputs (MacKay, 1992) and that the Boltzmann machine learning algorithm (Ackley, Hinton, & Sejnowski, 1985; Movellan & McClelland, 1993) can learn to produce correct conditional distributions of outputs given inputs. The original algorithm was very slow but recent variants are more efficient (Hinton & Salakhutdinov, 2006), and have been effectively used to model, for example, human numerosity judgments (Stoianov & Zorzi, 2012). (See Chapter 3 in this handbook for a fuller discussion.)

## 2.3 Three Foundational Models

This section outlines three of the landmark models in the emergence of connectionist theories of cognition. The models serve to illustrate the key principles of connectionism and demonstrate how these principles are relevant to explaining behavior in ways that are different from other prior approaches. The contribution of these models was twofold: they were better suited than alternative approaches to capturing the actual characteristics of human cognition, usually on the basis of their context-sensitive processing properties; and compared to existing accounts, they offered a sharper set of tools to drive theoretical progress and to stimulate empirical data collection. Each of these models significantly advanced its field.

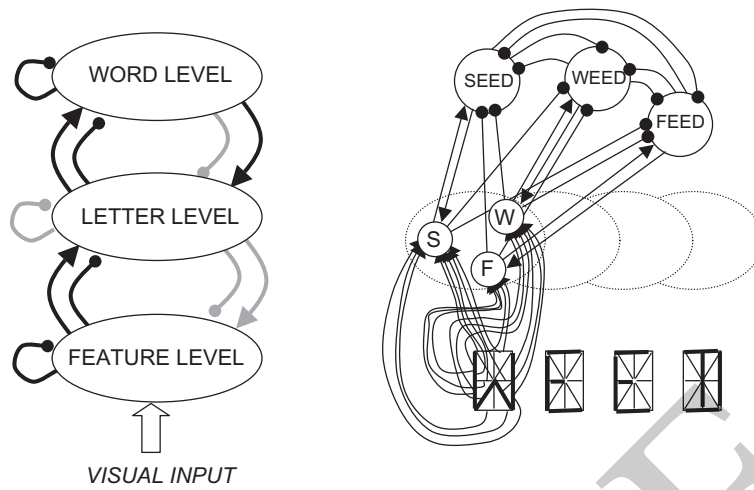
### 2.3.1 An Interactive Activation Model of Context Effects in Letter Perception (McClelland & Rumelhart, 1981, 1982)

The interactive activation model of letter perception illustrates two interrelated ideas. The first is that connectionist models naturally capture a graded constraint satisfaction process in which the influences of many different types of information are simultaneously integrated in determining, for example, the identity of a letter in a word. The second idea is that the computation of a perceptual representation of the current input (in this case, a word) involves the simultaneous and mutual influence of representations at *multiple levels of abstraction* – this is a core idea of parallel distributed processing.

The interactive activation model addressed itself to a puzzle in word recognition. By the late 1970s, it had long been known that people were better at recognizing letters presented in words than letters presented in random letter sequences. Reicher (1969) demonstrated that this was not the result of tending to guess letters that would make letter strings into words. He presented target letters either in words, unpronounceable nonwords, or on their own. The stimuli were then followed by a pattern mask, after which participants were presented with a forced choice between two letters in a given position. Importantly, both alternatives were equally plausible. Thus, the participant might be presented with WOOD and asked whether the third letter was O or R. As expected, forced-choice performance was more accurate for letters in words than for letters in nonwords or presented on their own. Moreover, the benefit of surrounding context was also conferred by pronounceable pseudowords (e.g., recognizing the P in SPET) compared to random letter strings, suggesting that subjects were able to bring to bear rules regarding the orthographic legality of letter strings during recognition.

Rumelhart and McClelland took the contextual advantage of words and pseudowords on letter recognition to indicate the operation of *top-down* processing. Previous theories had put forward the idea that letter and word recognition might be construed in terms of detectors which collect evidence consistent with the presence of their assigned letter or word in the input (Morton, 1969; Selfridge, 1959). Influenced by these theories, Rumelhart and McClelland built a computational simulation in which the perception of letters resulted from excitatory and inhibitory interactions of detectors for visual features. Importantly, the detectors were organized into different layers for letter features, letters and words, and detectors could influence each other both in a bottom-up and a top-down manner.

Figure 2.2 illustrates the structure of the Interactive Activation (IA) model, both at the macro level (left) and for a small section of the model at a finer level (right). The explicit motivation for the structure of the IA was neural: “[We] have adopted the approach of formulating the model in terms similar to the way in which such a process might actually be carried out in a neural or neural-like system” (McClelland & Rumelhart, 1981, p. 387). There were three main



**Figure 2.2** Interactive Activation model of context effects in letter recognition (McClelland & Rumelhart, 1981, 1982). Pointed arrows are excitatory connections, circular headed arrows are inhibitory connections. Left: macro view (connections in gray were set to zero in the implemented model). Right: micro view for the connections from the feature level to the first letter position for the letters S, W, and F (only excitatory connections shown) and from the first letter position to the word units SEED, WEED, and FEED (all connections shown).

assumptions of the IA model: (1) perceptual processing takes place in a system in which there are several levels of processing, each of which forms a representation of the input at a different level of abstraction; (2) visual perception involves parallel processing, both of the four letters in each word and of all levels of abstraction simultaneously; (3) perception is an interactive process in which conceptually driven and data-driven processing provide multiple, simultaneously acting constraints that combine to determine what is perceived.

The activation states of the system were simulated by a sequence of discrete time steps. Each unit combined its activation on the previous time step, its excitatory influences, its inhibitory influences, and a decay factor to determine its activation on the next time step. Connectivity was set at unitary values and along the following principles: in each layer, mutually exclusive alternatives should inhibit each other. For each unit in a layer, it excited all units with which it was consistent and inhibited all those with which it was inconsistent in layer immediately above. Thus in Figure 2.2, the first-position W letter unit has an excitatory connection to the WEED word unit but an inhibitory connection to the SEED and FEED word units. Similarly, a unit excited all units with which it was consistent and inhibited all those with which it was inconsistent in the layer immediately below. However, in the final implementation, top-down word-to-letter inhibition and within-layer letter-to-letter inhibition were set to zero (gray arrows, Figure 2.2).

The model was constructed to recognize letters in four-letter strings. The full set of possible letters was duplicated for each letter position, and a set of 1,179 word units created to represent the corpus of four-letter words. Word units were given base rate activation states at the beginning of processing to reflect their different frequencies. A trial began by clamping the feature units to the appropriate states to represent a letter string, and then observing the dynamic change in activation through the network. Conditions were included to allow the simulation of stimulus masking and degraded stimulus quality. Finally, a probabilistic response mechanism was added to generate responses from the letter level, based on the relative activation states of the letter pool in each position.

The model successfully captured the greater accuracy of letter detection for letters appearing in words and pseudowords compared to random strings or in isolation. Moreover, it simulated a variety of empirical findings on the effect of masking and stimulus quality, and of changing the timing of the availability of context. The results on the contextual effects of pseudowords are particularly interesting, since the model only contains word units and letter units and has no explicit representation of orthographic rules. Let us say on a given trial, the subject is required to recognize the second letter in the string SPET. In this case, the string will produce bottom-up excitation of the word units for SPAT, SPIT, and SPOT, which each share three letters. In turn, the word units will propagate top-down activation reinforcing activation of the letter P and so facilitating its recognition. Were this letter to be presented in the string XPQJ, no word units could offer similar top-down activation, hence the relative facilitation of the pseudoword. Interestingly, although these top-down “gang” effects produced facilitation of letters contained in orthographically legal nonword strings, the model demonstrated that they also produced facilitation in orthographically illegal, unpronounceable letter strings such as SPCT. Here, the same gang of SPAT, SPIT, and SPOT produce top-down support. Rumelhart and McClelland (1982) reported empirical support for this novel prediction. Therefore, although the model behaved *as if it contained orthographic rules driving recognition*, it did not in fact do so, because continued contextual facilitation could be demonstrated for strings that had gang support but violated the orthographic rules.

There are two specific points to note regarding the IA model. First, this early connectionist model was not adaptive – connectivity was set by hand. While the model’s behavior was shaped by the statistical properties of the language it processed, these properties were built into the structure of the system, in terms of the frequency of occurrence of letters and letter combinations in the words. Second, the idea of bottom-up excitation followed by competition amongst mutually exclusive possibilities is a strategy familiar in Bayesian approaches to cognition. In that sense, the IA bears similarity to more recent probability theory-based approaches to perception.

Subsequent work saw the principles of the IA model extended to the recognition of spoken words (the TRACE model: McClelland & Elman, 1986) and to



bilingual speakers where two languages must be incorporated in a single representational system (Grainger, Midgley & Holcomb, 2010; Thomas & van Heuven, 2005). The architecture was applied to other domains where multiple constraints were thought to operate during perception, for example in face recognition (Burton, Bruce, & Johnston, 1990). Within language, more complex architectures tried to recast the principles of the IA model in developmental settings, such as Plaut and Kello's (1999) model of the emergence of phonology from the interplay of speech comprehension and production.

The more general lesson to draw from the interactive activation model is the demonstration of multiple influences (feature, letter, and word-level knowledge) working simultaneously and in parallel to shape the response of the system; and the somewhat surprising finding that a massively parallel constraint satisfaction process of this form can appear to behave as if it contains rules (in this case, orthographic) when no such rules are included in the processing structure. At the time, the model brought into question whether it was necessary to postulate rules as processing structures to explain regularities in human behavior. This skepticism was brought into sharper focus by the next example.

### **2.3.2 On Learning the Past Tense of English Verbs (Rumelhart & McClelland, 1986)**

Rumelhart and McClelland's (1986) model of English past tense formation marked the real emergence of the PDP framework. Where the IA model used localist coding, the past tense model employed distributed coding. Where the IA model had handwired connection weights, the past tense model learned its weights via repeated exposure to a problem domain. However, the models share two common themes. Once more, the behavior of the past model will be driven by the statistics of the problem domain, albeit these will be carved into the model by training rather than sculpted by the modelers. Perhaps more importantly, there is a return to the idea that a connectionist system can exhibit rule-following behavior without containing rules as causal processing structures; but in this case, the rule-following behavior will be the product of learning and will accommodate a proportion of exception patterns that do not follow the general rule. The key point that the past tense model illustrates is how (approximate) conformity to the regularities of language – and even a tendency to produce new regular forms (e.g., regularizations like “thought” or past tenses for novel verbs like “wugged”) – can arise in a connectionist network without an explicit representation of a linguistic rule.

The English past tense is characterized by a predominant regularity in which the majority of verbs form their past tenses by the addition of one of three allomorphs of the “-ed” suffix to the base stem (walk/walked, end/ended, chase/chased). However, there is a small but significant group of verbs which form their past tense in different ways, including changing internal vowels (swim/swam), changing word final consonants (build/built), changing both internal vowels and final consonants (think/thought), an arbitrary relation of stem to

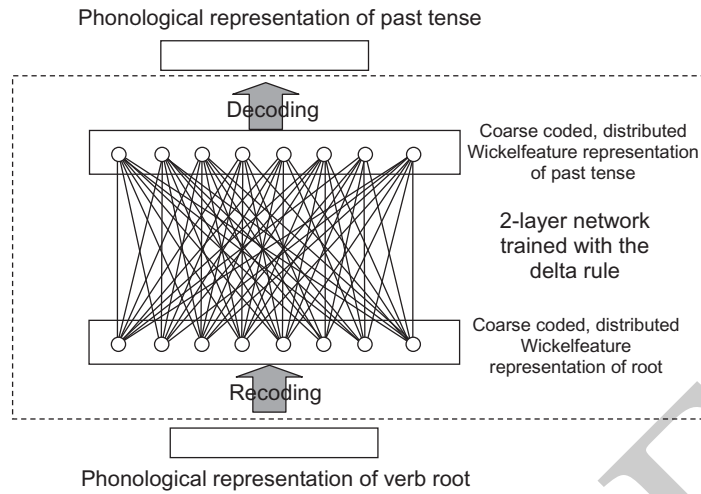
past tense (go/went), and verbs which have a past tense form identical to the stem (hit/hit). These so-called irregular verbs often come in small groups sharing a family resemblance (sleep/slept, creep/crept, leap/leapt) and usually have high token frequencies (see Pinker, 1999, for further details).

During the acquisition of the English past tense, children show a characteristic U-shaped developmental profile at different times for individual irregular verbs. Initially they use the correct past tense of a small number of high frequency regular and irregular verbs. Latterly, they sometimes produce “overregularized” past tense forms for a small fraction of their irregular verbs (e.g., *thought*) (Marcus, Pinker, Ullman, Hollander, Rosen, & Xu, 1992), along with other, less frequent errors (Xu & Pinker, 1995). They are also able to extend the past tense “rule” to novel verbs (e.g., *wug* – *wugged*). Finally, in older children, performance approaches ceiling on both regular and irregular verbs (Berko, 1958; Ervin, 1964; Kuczaj, 1977).

In the early 1980s, it was held that this pattern of behavior represented the operation of two developmental mechanisms (Pinker, 1984). One of these was symbolic and served to learn the regular past tense “rule,” while the other was associative and served to learn the exceptions to the rule. The extended phase of overregularization errors corresponded to difficulties in integrating the two mechanisms, specifically a failure of the associative mechanism to block the function of the symbolic mechanism. That the child comes to the language acquisition situation armed with these two mechanisms (one of them full of blank rules) was an *a priori* commitment of the developmental theory.

By contrast, Rumelhart and McClelland (1986) proposed that a single network that does not distinguish between regular and irregular past tenses is sufficient to learn past tense formation. The architecture of their model is shown in Figure 2.3. A phoneme-based representation of the verb root was recoded into a more distributed, coarser (more blurred) format, which they called “Wickelfeatures.” The stated aim of this recoding was to produce a representation that (a) permitted differentiation of all of the root forms of English and their past tenses, and (b) provided a natural basis for generalizations to emerge about what aspects of a present tense correspond to what aspects of a past tense. This format involved representing verbs over 460 processing units. A two-layer network was then used to associate the Wickelfeature representations of the verb root and past tense form. A final decoding network was then used to derive the closest phoneme-based rendition of the past tense form and reveal the model’s response (the decoding part of the model was somewhat restricted by computer processing limitations of the machines available at the time).

The connection weights in the two-layer network were initially randomized. The model was then trained in three phases, in each case using the delta rule to update the connection weights after each verb root/past tense pair was presented (see Section 2.1.2). In Phase 1, the network was trained on ten high frequency verbs, two regular and eight irregular, in line with the greater proportion of irregular verbs amongst the most frequent verbs in English. Phase 1 lasted for ten presentations of the full training set (or “epochs”). In Phase 2,



**Figure 2.3** Two-layer network for learning the mapping between the verb roots and past tense forms of English verbs (Rumelhart & McClelland, 1986). Phonological representations of verbs are initially encoded into a coarse, distributed “Wickelfeature” representation. Past tenses are decoded from the Wickelfeature representation back to the phonological form. Later connectionist models replaced the dotted area with a three-layer feedforward backpropagation network (e.g., Plunkett & Marchman, 1991, 1993).

the network was trained on 410 medium frequency verbs, 334 regular and 76 irregular, for a further 190 epochs. In Phase 3, no further training took place, but 86 lower frequency verbs were presented to the network to test its ability to generalize its knowledge of the past tense domain to novel verbs.

There were four key results for this model. First, it succeeded in learning both regular and irregular past tense mappings in a single network that made no reference to the distinction between regular and irregular verbs. Second, it captured the overall pattern of faster acquisition for regular verbs than irregular verbs, a predominant feature of children’s past tense acquisition. Third, the model captured the U-shaped profile of development: an early phase of accurate performance on a small set of regular and irregular verbs, followed by a phase of overregularization of the irregular forms, and finally recovery for the irregular verbs and performance approaching ceiling on both verb types. Fourth, when the model was presented with the low-frequency verbs on which it had not been trained, it was able to generalize the past tense rule to a substantial proportion of them, as if it had indeed learned a rule. Additionally, the model captured more fine-grained developmental patterns for subsets of regular and irregular verbs, and generated several novel predictions.

Rumelhart and McClelland explained the generalization abilities of the network in terms of the *superpositional* memory of the two-layer network. All the associations between the distributed encodings of verb root and past tense forms must be stored across the single matrix of connection weights. As a result, similar patterns blend into one another and reinforce each other. Generalization

is contingent on the similarity of verbs at input. Were the verbs to be presented using an orthogonal, localist scheme (e.g., 420 units, one per verb), then there would be no similarity between the verbs, no blending of mappings, no generalization, and therefore no regularization of novel verbs. As the authors state, “it is the statistical relationships among the base forms themselves that determine the pattern of responding. The network merely reflects the statistics of the featural representations of the verb forms” (p. 267). Based on the model’s successful simulation of the profile of language development in this domain and, compared to the dual mechanism model, its more parsimonious *a priori* commitments, Rumelhart and McClelland viewed their work on past tense morphology as a step towards a revised understanding of language knowledge, language acquisition, and linguistic information processing in general.

The past tense model stimulated a great deal of subsequent debate, not least because of its profound implications for theories of language development (no rules!). The model was initially subjected to concentrated criticism. Some of this was overstated – for instance, the use of domain-general learning *principles* (such as distributed representation, parallel processing, and the delta rule) to acquire the past tense in a single network was interpreted as a claim that all of language acquisition could be captured by the operation of a single domain-general learning *mechanism*. Such an absurd claim could be summarily dismissed. However, as it stood, the model made no such claim: its generality was in the processing principles. The model itself represented a domain-specific system dedicated to learning a small part of language. Nevertheless, a number of the criticisms were more telling: the Wickelfeature representational format was not psycholinguistically realistic; the generalization performance of the model was relatively poor; the U-shaped developmental profile appeared to be a result of abrupt changes in the composition of the training set; and the actual response of the model was hard to discern because of problems in decoding the Wickelfeature output into a phoneme string (Pinker & Prince, 1988).

The criticisms and following rejoinders were interesting in a number of ways. First, there was a stark contrast between the precise, computationally implemented connectionist model of past tense formation and the verbally specified two-system theory (e.g., Marcus, Pinker, Ullman, Hollander, Rosen, & Xu, 1992). The implementation made simplifications but was readily evaluated against quantitative behavioral evidence; it made predictions and it could be falsified. The verbal theory by contrast was vague – it was hard to know how or whether it would work or exactly what behaviors it predicted (Thomas, Forrester, & Richardson, 2006). Therefore, it could only be evaluated on loose qualitative grounds. Second, the model stimulated a great deal of new multi-disciplinary research in the area. Today, inflectional morphology (of which past tense is a part) is one of the most studied aspects of language processing in children, in adults, in second language learners, in adults with acquired brain damage, in children and adults with neurogenetic disorders, and in children with language impairments, using psycholinguistic methods, event-related potential measures of brain activity, functional magnetic resonance imaging,

and behavioral genetics . . . This rush of science illustrates the essential role of computational modeling in driving forward theories of human cognition. Third, further modifications and improvements to the past tense model have highlighted how researchers go about the difficult task of understanding which parts of their model represent the key theoretical claims and which are implementational details. Simplification is inherent to modeling but successful modeling relies on making the *right* simplifications to focus on the process of interest. For example, in subsequent models, the Wickelfeature representation was replaced by more plausible phonemic representations based on articulatory features; the recoding/two-layer-network/decoding component of the network (the dotted rectangle in Figure 2.3) that was trained with the delta rule was replaced by a three-layer feedforward network trained with the backpropagation algorithm; and the U-shaped developmental profile was demonstrated in connectionist networks trained with a smoothly growing training set of verbs or even with a fixed set of verbs (see, e.g., Plunkett & Marchman, 1991, 1993, 1996).

The English past tense model prompted further work within inflectional morphology in other languages (pluralization in German: Goebel & Indefrey, 2000; pluralization in Arabic: Plunkett & Nakisa, 1997), as well as models that explored the possible causes of deficits in acquired and developmental disorders such as aphasia, developmental language disorder, and Williams syndrome (e.g., Hoeffner & McClelland, 1993; Joanisse & Seidenberg, 1999; Thomas & Karmiloff-Smith, 2003a; Thomas & Knowland, 2014). More recent work treats the past tense as one role of a more general system which has the goal of outputting the phonological form of words appropriate to the syntactic context of the sentence in which they appear – whether this involves the tense of verbs, the number of nouns, or the comparative of adjectives (Karaminis & Thomas, 2010, 2014). Moreover, the idea that rule-following behavior could emerge in a developing system that also has to accommodate exceptions to the rules was also successfully pursued via connectionist modeling in the domain of reading (e.g., Plaut et al., 1996). This led to work that also considered various forms of acquired and developmental dyslexia.

For the past tense itself, there remains much interest in the topic as a crucible to test theories of language development. There is now extensive evidence from child development, adult cognitive neuropsychology, developmental neuropsychology, and functional brain imaging to suggest partial dissociations between performance on regular and irregular inflection under various conditions. For the connectionist approach, the dissociations represent the integration of multiple information sources, syntactic, lexical semantic, and phonological. Regular and irregular inflections depend differently on these sources depending on statistical properties of the mappings, explaining the dissociations. For the two-system approach, the dissociations represent separate contributions of causal rules and associative memory. (See Pater, 2019, and Kirov & Cotterell, 2018, for more recent reviews of this debate from the perspective of linguistics). Nevertheless, the force of the original past tense model remains: so long as there are regularities in the statistical structure of a

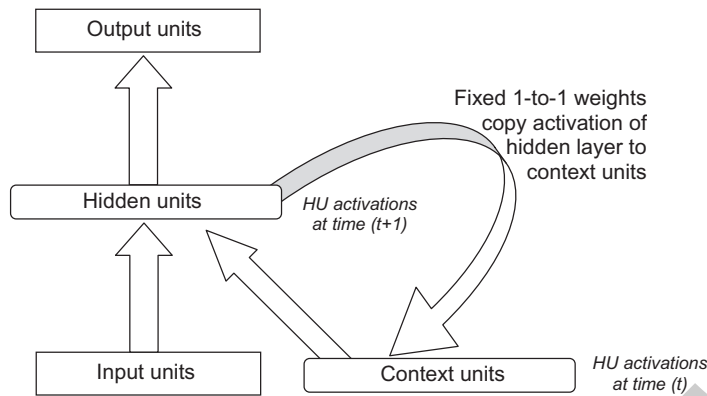
problem domain, a massively parallel constraint satisfaction system can learn these regularities and extend them to novel situations. Moreover, as with humans, the behavior of the system is flexible and context sensitive – it can accommodate regularities and exceptions within a single processing structure.

### 2.3.3 Finding Structure in Time (Elman, 1990)

This section introduces the notion of the simple recurrent network and its application to language. As with past tense, the key point of the model will be to show how conformity to regularities of language can arise without an explicit representation of a linguistic rule. Moreover, the following simulations will demonstrate how learning can lead to the discovery of useful internal representations that capture conceptual and linguistic structure on the basis of the cooccurrences of words in sentences.

The IA model exemplified connectionism's commitment to parallelism: all of the letters of the word presented to the network were recognized in parallel and processing occurred simultaneously at different levels of abstraction. But not all processing can be carried out in this way. Some human behaviors intrinsically revolve around temporal sequences. Language, action planning, goal-directed behavior, and reasoning about causality are examples of domains that rely on events occurring in sequences. How has connectionism addressed the processing of temporally unfolding events? One solution was offered in the TRACE model of spoken word recognition (McClelland & Elman, 1986) where a word was specified as a sequence of phonemes. In that case, the architecture of the system was duplicated for each time slice and the duplicates wired together. This allowed constraints to operate over items in the sequence to influence recognition. In other models, a related approach was used to convert a temporally extended representation into a spatially extended one. For example, in the past tense model, all the phonemes of a verb were presented across the input layer. This could be viewed as a sequence if one assumed that the representation of the first phoneme represents time slice  $t$ , the representation of the second phoneme represents time slice  $t+1$ , and so on. As part of a comprehension system, this approach assumes a buffer that can take sequences and convert them to a spatial vector. However, this solution is fairly limited, as it necessarily precommits to the size of the sequences that can be processed at once (i.e., the size of the input layer).

Elman (1990, 1991) offered an alternative and more flexible approach to processing sequences, proposing an architecture that has been extremely influential and much used since. Elman drew on the work of Jordan (1986) who had proposed a model that could learn to associate a "plan" (i.e., a single input vector) with a series of "actions" (i.e., a sequence of output vectors). Jordan's model contained recurrent connections permitting the hidden units to "see" the network's previous output (via a set of "state" input units that are given a copy of the output on the previous time step). The facility for the network to shape its next output according to its previous response constitutes a kind of memory.



**Figure 2.4** Elman's simple recurrent network architecture for finding structure in time (Elman, 1991, 1993). Connections between input and hidden, context and hidden, and hidden and output layers are trainable. Sequences are applied to the network element by element in discrete time steps; the context layer contains a copy of the hidden unit activations on the previous time step transmitted by fixed, one-to-one connections.

Elman's innovation was to build a recurrent facility into the internal units of the network, allowing it to compute statistical relationships across sequences of inputs and outputs. To achieve this, first time is discretized into a number of slices. On time step  $t$ , an input is presented to the network and causes a pattern of activation on hidden and output layers. On time step  $t + 1$ , the next input in the sequence of events is presented to the network. However, crucially, a copy of the activation of the hidden units on time step  $t$  is transmitted to a set of internal "context" units. This activation vector is also fed to the hidden units on time step  $t + 1$ . Figure 2.4 shows the architecture, known as the *simple recurrent network* (SRN). It is usually trained with the backpropagation algorithm (see Section 2.2.3) as a multi-layer feedforward network, ignoring the origin of the information on the context layer.

Each input to the SRN is therefore processed in the context of what came before, but in a way subtly more powerful than the Jordan network. The input at  $t + 1$  is processed in the context of the activity produced on the hidden units by the input at time  $t$ . Now consider the next time step. The input at time  $t + 2$  will be processed along with activity from the context layer that is shaped by *two* influences:

(the input at  $t + 1$  (shaped by the input at  $t$ ))

The input at time  $t + 3$  will be processed along with activity from the context layer that is shaped by *three* influences:

(the input at  $t + 2$  (shaped by the input at  $t + 1$  (shaped by the input at  $t$ )))

The recursive flavor of the information contained in the context layer means that each new input is processed in the context of the *full history* of previous

inputs. This permits the network to learn statistical relationships across sequences of inputs or, in other words, to find structure in time.

In his original paper of 1990, Elman demonstrated the powerful properties of the SRN with two examples. In the first, the network was presented with a sequence of letters made up of concatenated words, e.g.:

### **Many Years Ago a Boy and Girl Lived by the Sea they Played Happily**



Each letter was represented by a distributed binary code over five input units. The network was trained to predict the next letter in the sentence for 200 sentences constructed from a lexicon of fifteen words. There were 1,270 words and 4,963 letters. Since each word appeared in many sentences, the network was not particularly successful at predicting the next letter when it got to the end of each word, but within a word it was able to predict the sequences of letters. Using the accuracy of prediction as a measure, one could therefore identify which sequences in the letter string were words: they were the sequences of good prediction bounded by high prediction errors. The ability to extract words was of course subject to the ambiguities inherent in the training set (e.g., for *the* and *they*, there is ambiguity after the third letter). Elman suggested that if the letter strings are taken to be analogous to the speech sounds available to the infant, the SRN demonstrates a possible mechanism to extract words from the continuous stream of sound that is present in infant-directed speech. Elman's work contributed to the increasing interest in the statistical learning abilities of young children in language and cognitive development (e.g., Saffran & Kirkham, 2018; Saffran, Newport, & Aslin, 1996).

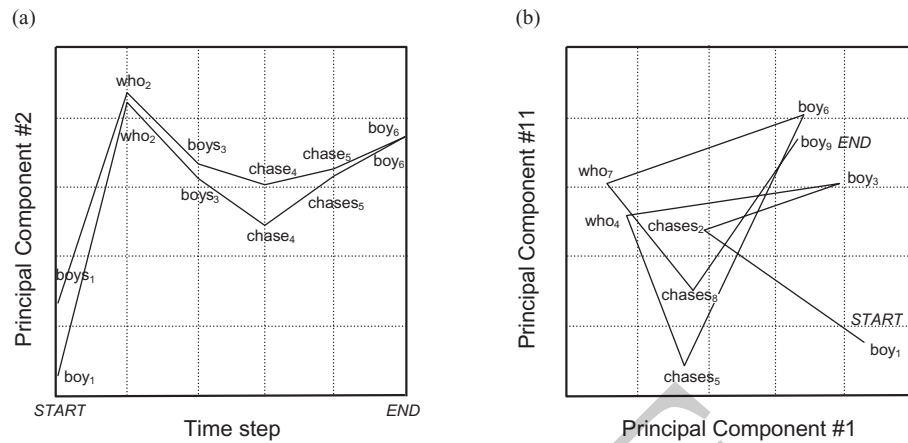
In the second example, Elman created a set of 10,000 sentences by combining a lexicon of twenty-nine words and a set of short sentence frames (noun + [transitive] verb + noun; noun + [intransitive] verb). There was a separate input and output unit for each word and the SRN was trained to predict the next word in the sentence. During training, the network's output came to approximate the transitional probabilities between the words in the sentences – that is, it could predict the next word in the sentences as much as this was possible. Following the first noun, the verb units would be more active as the possible next word, and verbs that tended to be associated with this particular noun would be more active than those that did not. At this point, Elman examined the similarity structure of the internal representations to discover how the network was achieving its prediction ability. He found that the internal representations were sensitive to the difference between nouns and verbs, and within verbs, to the difference between transitive and intransitive verbs. Moreover, the network was also sensitive to a range of semantic distinctions: not only were the internal states induced by nouns split into animate and inanimate, but the pattern for “woman” was most similar to “girl,” and that for “man” was most similar to “boy.” The network had learnt



to structure its internal representations according to a mix of syntactic and semantic information because these information states were the best way to predict how sentences would unfold. Elman concluded that the representations induced by connectionist networks need not be flat but could include hierarchical encodings of category structure.

Based on his finding, Elman also argued that the SRN was able to induce representations of entities that varied according to their context of use. This contrasts with classical symbolic representations that retain their identity irrespective of the combinations into which they are put, a property called “compositionality.” This claim is perhaps better illustrated by a second paper Elman published two years later called “The importance of starting small” (1993). In this later paper, Elman explored whether rule-based mechanisms are required to explain certain aspects of language performance, such as syntax. He focused on “long-range dependencies,” which are links between words that depend only on their syntactic relationship in the sentence and, importantly, not on their separation in a sequence of words. For example, in English, the subject and main verb of a sentence must agree in number. If the noun is singular, so must be the verb; if the noun is plural, so must be the verb. Thus, in the sentence “The **boy chases** the cat,” *boy* and *chases* must both be singular. But this is also true in the sentence “The **boy** whom the boys chase **chases** the cat.” In the second sentence, the subject and verb are further apart in the sequence of words, but their relationship is the same; moreover, the words are now separated by plural tokens of the same lexical items. Rule-based representations of syntax were thought to be necessary to encode these long-distance relationships because, through the recursive nature of syntax, the words that have to agree in a sentence can be arbitrarily far apart.

Using an SRN trained on the same prediction task as that outlined above but now with more complex sentences, Elman (1993) demonstrated that the network was able to learn these long-range dependencies even across the separation of multiple phrases. If *boy* was the subject of the sentence, when the network came to predict the main verb *chase* as the next word, it predicted that it should be in the singular. The method by which the network achieved this ability is of particular interest. Once more, Elman explored the similarity structure in the hidden unit representations, using principal component analyses to identify the salient dimensions of similarity across which activation states were varying. This enabled him to reduce the high dimensionality of the internal states (150 hidden units were used) to a manageable number in order to visualize processing. Elman was then able to plot the *trajectories* of activation as the network altered its internal state in response to each subsequent input. Figure 2.5 depicts these trajectories as the network processes different multi-phrase sentences, plotted with reference to particular dimensions of principal component space. This figure demonstrates that the network adopted similar states in response to particular lexical items (e.g., tokens of *boy*, *who*, *chases*), but that it modified the pattern slightly according to the grammatical status of the word. In Figure 2.5a, the second principal component appears to encode



**Figure 2.5** Trajectory of internal activation states as the SRN processes sentences (Elman, 1993). The data show positions according to the dimensions of a principal components analysis (PCA) carried out on hidden unit activations for the whole training set. Words are indexed by their position in the sequence but represent activation of the same input unit for each word. (a) PCA values for the second principal component as the SRN processes two sentences, “Boy who boys chase chases boy” or “Boys who boys chase chase boy”; (b) PCA values for the first and eleventh principal components as the SRN processes “Boy chases boy who chases boy who chases boy.”

singularity/plurality. Figure 2.5b traces the network’s state as it processes two embedded relative clauses containing iterations of the same words. Each clause exhibits a related but slightly shifted triangular trajectory to encode its role in the syntactic structure.

The importance of this model is that it prompts a different way to understand the processing of sentences. Previously one would view symbols as possessing fixed identities and as being bound into particular grammatical roles via a syntactic construction. In the connectionist system, sentences are represented by trajectories through activation space in which the activation pattern for each word is subtly shifted according to the context of its usage. The implication is that the property of compositionality at the heart of the classical symbolic computational approach may not be necessary to process language.

Elman (1993) also used this model to investigate a possible advantage to learning that could be gained by initially restricting the complexity of the training set. At the start of training, the network had its memory reset (its context layer wiped) after every third or fourth word. This window was then increased in stages up to six to seven words across training. The manipulation was intended to capture maturational changes in working memory in children. Elman (1993) reported that *starting small* enhanced learning by allowing the network to build simpler internal representations that were later useful for unpacking the structure of more complex sentences (see Rohde & Plaut, 1999, for discussion and further simulations). This idea resonated with developmental

psychologists in its demonstration of the way in which learning and maturation might interact in constructing cognition (Elman et al., 1996).

Recurrent models were subsequently extended to consider other domains where temporal information about sequence is important. For example, Botvinick and Plaut (2004) demonstrated how simple recurrent networks can capture the control of routine sequences of actions without the need for schema hierarchies. Elman and McRae (2019) used simple recurrence to construct a model of semantic event knowledge, that is, what tends to happen in different situations involving actors and agents. The model learned both the internal structure of activities as well as the temporal structure that organizes activity sequences. Cleeremans and colleagues demonstrated how simple recurrent models were a useful architecture to understand phenomena within implicit learning, which often involve detecting patterns within sequences of stimuli (see Cleeremans & Dienes, 2008).

In the domain of language processing, meanwhile, subsequent progress was initially slow (Christiansen & Chater, 2001). The ability of simple recurrent networks to induce structured representations containing grammatical and semantic information from word sequences prompted the view that associative statistical learning mechanisms might play a much more central role in language acquisition. This innovation was especially welcome given that symbolic theories of sentence processing do not offer a ready account of language development. Indeed, they are largely identified with the nativist view that little in syntax develops. But a limitation of Elman's initial simulations was that the prediction task does not learn any categorizations over the input set. While the simulations demonstrate that information important for language comprehension and production can be induced from word sequences, neither task was performed.

Recurrent neural network approaches to sentence processing have gone in two directions. In terms of cognitive modeling, connectionist simulations have included more differentiated structure to learn mappings between messages and word sequences, including limited use of binding to temporarily link concepts and roles (Chang, Dell, & Bock, 2006). ~~Most recently,~~ the model has been applied to how children learn the relationship between declarative (statement) and interrogative (question) sentences (Fitz & Chang, 2017). In terms of engineering approaches, deep recurrent neural networks have been scaled up to an extent where they can achieve automatic translation between sentences in different languages with a reasonable degree of accuracy, such as in the case of Google Translate (Wu et al., 2016). The architecture of Google Translate includes a deep recurrent neural network (eight layers) that encodes a sentence of the first language in a vector of numbers, and a decoder network (also eight layers) that learns to map to a similar vector in the second language and then to an output sequence. The mapping between encoder and decoder is mediated by an "attention" mechanism that gives flexibility on which parts of the first sentence might map to which parts of the second sentence. The overall system is trained to map between millions of sentences in the two languages.

While the degree of accuracy of translation is unimaginable from the perspective of the early PDP models and must rely heavily on the syntactic information in the respective languages, from a cognitive perspective, it contains no representation of sentence meaning. The shallowness of the mapping between languages becomes apparent when real world knowledge is required to solve ambiguities in sentence processing, such as which pronouns refer to which nouns; here, Google Translate can perform poorly (Hofstadter, 2018). However, within linguistics, the successes of machine translation by deep recurrent neural networks has focused attention on learning theory to constrain theories of grammar (Pater, 2019). Moreover, the new recurrent network translation models lend credence to early claims by PDP researchers (e.g., Rumelhart, Smolensky, McClelland, & Hinton, 1986) that thoughts – although they can be expressed as sentences – are represented in the brain as vectors (patterns of neural activation) and that reasoning is a sequence of transitions between such vectors. As of mid 2020, further breakthroughs in machine language processing have occurred (Brown et al., 2020). The latest models now resolve referential ambiguities better than earlier versions, and their internal representations appear to capture syntactic structure in language better than critics expected (Manning et al., 2020). However, they still fail at capturing human understanding of common-sense physical relationships, indicating they are still somewhat shallow language processors. An exciting next step for neural language models will be to place them within systems that understand and communicate about real or hypothetical situations, since ultimately this is what language is for (McClelland et al., 2020).

In sum, then, Elman's work demonstrates how simple connectionist architectures can learn statistical regularities over temporal sequences. These systems may indeed be sufficient to produce many of the behaviors that linguists have described with grammatical rules. However, in the connectionist system, the underlying primitives are context-sensitive representations of words and trajectories of activation through recurrent circuits. Such representations appear to be playing a more and more important role in theories of how humans process – and even understand – natural language.

## 2.4 Connectionist Influences on Cognitive Theory

Connectionism offers an *explanation* of human cognition because instances of behavior in particular cognitive domains can be explained with respect to a set of general principles (parallel distributed processing) and the conditions of the specific domains. However, from the accumulation of successful models, it is also possible to discern a wider influence of connectionism on the nature of theorizing about cognition, and this is perhaps a truer reflection of its impact. How has connectionism made us think differently about cognition?

### 2.4.1 Knowledge versus Processing

One area where connectionism has changed the basic nature of theorizing is memory. According to the old model of memory based on the classical computational metaphor, the information in long-term memory (e.g., on the hard disk) has to be moved into working memory (the CPU) for it to be operated on, and the long-term memories are laid down via a domain-general buffer of short-term memory (RAM). In this type of system, then, long-term memory is separated from processing. It is relatively easy to shift informational content between different systems, back and forth between central processing and short- and long-term stores. Computation is predicated on variables: the same binary string can readily be instantiated in different memory registers or encoded onto a permanent medium.

By contrast, knowledge is hard to move about in connectionist networks because it is encoded in the weights. For example, in the past tense model, knowledge of the past tense rule “add -ed” is distributed across the weight matrix of the connections between input and output layers. The difficulty in portability of knowledge is inherent in the principles of connectionism – Hebbian learning alters connection strengths to reinforce desirable activation states in connected units, tying knowledge to structure. If the foundational premise is that knowledge will be very difficult to move about in the human information processing system, what kind of cognitive architecture results? There are four main themes.

First, it is necessary to distinguish between two different ways in which knowledge can be encoded: *active* and *latent* representations (Munakata & McClelland, 2003). Latent knowledge corresponds to the information stored in the connection weights from accumulated experience. By contrast, active knowledge is information contained in the current activation states of the system. Clearly the two are related, since the activation states are constrained by the connection weights. But, particularly in recurrent networks, there can be subtle differences. Active states contain a trace of recent events (how things are at the moment) while latent knowledge represents a history of experience (how things tend to be). Differences in the ability to maintain the active states (e.g., in the strength of recurrent circuits) can produce errors in behavior where the system lapses into more typical ways of behaving (Morton & Munakata, 2002; Munakata, 1998).

Second, if information does need to be moved around the system, for example from a more instance-based (episodic) system to a more general (semantic) system, this will require special structures and special (potentially time consuming) processes. Thus McClelland, McNaughton, and O’Reilly (1995) proposed a dialogue between separate stores in the hippocampus and neocortex to gradually transfer knowledge from episodic to semantic memory (see O’Reilly, Bhattacharyya, Howard, & Ketza, 2014). For example, French, Ans, and Rousset (2001) proposed a special method to transfer knowledge

between two memory systems: internally generated noise produces “pseudopatterns” from one system that contain the central tendencies of its knowledge; the second memory system is then trained with this extracted knowledge to effect the transfer.

Third, information will be processed in the same substrate where it is stored. Therefore, long-term memories will be active structures and will perform computations on content. An external strategic control system plays the role of differentially activating the knowledge in this long-term system that is relevant to the current context. In anatomical terms, this distinction broadly corresponds to frontal/anterior (strategic control) and posterior (long-term) cortex, with posterior cortex comprising a suite of content-specific processing systems. The design means, somewhat counter-intuitively, that the control system has no content. Rather, the control system contains placeholders that serve to activate different regions of the long-term system. The control system may contain plans (sequences of placeholders) and it may be involved in learning abstract concepts (using a placeholder to temporarily co-activate previously unrelated portions of long-term knowledge while Hebbian learning builds an association between them) but it does not contain content in the sense of a domain-general working memory. The study of frontal systems then becomes an exploration of the activation dynamics of these placeholders and their involvement in learning (see, e.g., work by Botvinick & Cohen, 2014; Davelaar & Usher, 2002; Haarmann & Usher, 2001; O’Reilly, Braver, & Cohen, 1999; Usher & McClelland, 2001).

Similarly, connectionist research has explored how activity in the control system can be used to modulate the efficiency of processing elsewhere in the system, for instance to implement selective attention. For example, in an early model, Cohen, Dunbar, and McClelland (1990) demonstrated how task units could be used to differentially modulate word naming and color naming processing channels in a model of the color-word Stroop task. Here, latent knowledge interacted with the operation of task control, so that it was harder to selectively attend to color naming and ignore information from the more practiced word-naming channel than vice versa. This work was later extended to demonstrate how deficits in the strategic control system (prefrontal cortex) could lead to problems in selective attention in disorders like schizophrenia (see Botvinick & Cohen, 2014, for a review).

Lastly, the connectionist perspective on memory alters the conception of *domain generality* in processing systems. It is unlikely that there are any domain-general processing systems that serve as a “Jack of all trades,” i.e., that can move between representing the content of multiple domains. However, there may be domain-general systems that are involved in modulating many disparate processes without taking on the content of those systems, either via direct connectivity or through the regional modulation of neurotransmitter levels. This type of general system might be called one with “a finger in every pie.” Meanwhile, short-term or working memory (as exemplified by the active representations contained in the recurrent loop of a network) is likely to exist as

a devolved panoply of discrete systems, each with its own content-specific loop. For example, research in the neuropsychology of language tends to support the existence of separate working memories for phonological, semantic, and syntactic information (MacDonald & Christiansen, 2002). And one might expect recurrent loops in the prefrontal cortex to maintain information about current goal states and positions in task sequences. From a connectionist perspective, therefore, and in contrast to traditional cognitive theory, *there is no such thing as working memory as a general mechanism*; rather it is a content-specific activity carried out in multiple systems.

### 2.4.2 Cognitive Development

A key feature of PDP models is the use of a learning algorithm for modifying the patterns of connectivity as a function of experience. Compared to symbolic, rule-based computational models, this has made them a more sympathetic formalism for studying cognitive development (Elman et al., 1996). The combination of domain-general processing principles, domain-specific architectural constraints, and structured training environments has enabled connectionist models to give accounts of a range of developmental phenomena. These include infant category development, language acquisition and reasoning in children (see Mareschal & Thomas, 2007; see also Chapter 23 in this handbook).

Connectionism has become aligned with a resurgence of interest in statistical learning, and a more careful consideration of the information available in the child's environment that may feed their cognitive development. One central debate revolves around how children can become "cleverer" as they get older, appearing to progress through qualitatively different stages of reasoning. Connectionist modeling of the development of children's reasoning was able to demonstrate that continuous incremental changes in the weight matrix driven by algorithms such as backpropagation can result in nonlinear changes in surface behavior, suggesting that the stages apparent in behavior may not necessarily be reflected in changes in the underlying mechanism (McClelland, 1989). Other connectionists have argued that algorithms able to supplement the computational resources of the network as part of learning may also provide an explanation for the emergence of more complex forms of behavior with age in so-called constructivist networks (e.g., cascade correlation; see Shultz, 2003; see also Chapter 23 in this handbook).

The key contribution of connectionist models in the area of developmental psychology has been to specify detailed, implemented models of transition mechanisms that demonstrate how the child can move between producing different patterns of behavior. This was a crucial addition to a field that has accumulated vast amounts of empirical data cataloguing what children are able to do at different ages. The specification of mechanism is also important to counter some strongly empiricist views that simply to identify statistical information in the environment suffices as an explanation of development; instead, it is necessary to show how a mechanism could use this statistical information to

acquire some cognitive capacity. Moreover, when connectionist models are applied to development, it often becomes apparent that passive statistical structure is not the key factor; rather, the relevant statistics are in the transformation of the statistical structure of the environment to the output or the behavior that is relevant to the child, thereby appealing to notions like the regularity, consistency, and frequency of input–output mappings.

Connectionist approaches to development have influenced understanding of the nature of the knowledge that children acquire. For example, Mareschal et al. (2007) argued that many mental representations of knowledge are partial (i.e., capture only some task-relevant dimensions) and only some dimensions of knowledge may be activated in any given situation; the existence of explicit language may blind people to the fact that there could be a limited role for truly abstract knowledge in the normal operation of the cognitive system (Westermann et al., 2007; Westermann, Thomas, & Karmiloff-Smith, 2010).

One important topic area gaining more attention is the use of connectionist models to capture aspects of numerical and mathematical cognition. This is an attractive application area since it has now become clear that an understanding of exact number (Gordon, 2004), and even the precision of approximate number estimation (Piazza et al., 2013) are highly experience-dependent. Building on earlier work by Verguts and Fias (2004), Stoianov and Zorzi (2012) introduced a neural network that captured aspects of adult human numerical estimation abilities, and Tesolin, Zou, and McClelland (2020) applied a similar approach to capture experience-dependent developmental increases in precision. More recent work using newer neural network architectures captures the emergence of an understanding of the exact number system through experience with an ensemble of distinct but underlyingly overlapping exact-number dependent tasks (Sabatiel, McClelland, & Solstad, 2020).

### **2.4.3 The Study of Acquired Disorders in Cognitive Neuropsychology**

Traditional cognitive neuropsychology of the 1980s was predicated on the assumption of underlying modular structure, i.e., that the cognitive system comprises a set of independently functioning components. Patterns of selective cognitive impairment after acquired brain damage could then be used to construct models of normal cognitive function. The traditional models comprised box-and-arrow diagrams that sketched out rough versions of cognitive architecture, informed both by the patterns of possible selective deficit (which bits can fail independently) and by a task analysis of what the cognitive system probably has to do.

In the initial formulation of cognitive neuropsychology, caution was advised in attempting to infer cognitive architecture from behavioral deficits, since a given pattern of deficits might be consistent with a number of underlying architectures (Shallice, 1988). It is in this capacity that connectionist models have been extremely useful. They have both forced more detailed specification of proposed cognitive models via implementation and also permitted



assessment of the range of deficits that can be generated by damaging these models in various ways. For example, models of reading have demonstrated that the ability to decode written words into spoken words and recover their meanings can be learned in a connectionist network; and when this network is damaged by, say, lesioning connection weights or removing hidden units, various patterns of acquired dyslexia can be simulated (e.g., Plaut et al., 1996; Woollams, 2014). Connectionist models of acquired deficits have grown to be an influential aspect of cognitive neuropsychology and have been applied to domains such as language, memory, semantics, and vision (see Cohen, Johnstone, & Plunkett, 2000, for examples).

Several ideas have gained their first or clearest grounding via connectionist modeling. One of these ideas is that patterns of breakdown can arise from the statistics of the problem space (i.e., the mapping between input and output) rather than from structural distinctions in the processing system. In particular, connectionist models have shed light on a principal inferential tool of cognitive neuropsychology, the *double dissociation*. The line of reasoning argues that if in one patient, ability A can be lost while ability B is intact, and in a second patient, ability B can be lost while ability A is intact, then the two abilities may be generated by independent underlying mechanisms. In a connectionist model of category-specific impairments of semantic memory, Devlin et al. (1997) demonstrated that a single undifferentiated network trained to produce two behaviors could show a double dissociation between them simply as a consequence of different levels of damage. This can arise because the mappings associated with the two behaviors lead them to have different sensitivity to damage. For a small level of damage, performance on A may fall off quickly while performance on B declines more slowly; for a high level of damage, A may be more robust than B. The reverse pattern of relative deficits implies nothing about structure.

Connectionist researchers have often set out to demonstrate that, more generally, double dissociation methodology is a flawed form of inference, on the grounds that such dissociations arise relatively easily from parallel distributed architectures where function is spread across the whole mechanism. However, on the whole, when connectionist models show robust double dissociations between two behaviors (for equivalent levels of damage applied to various parts of the network and over many replications), it does tend to be because different internal processing structures (units or layers or weights) or different parts of the input layer or different parts of the output layer are differentially important for driving the two behaviors – that is, there is specialization of function. Connectionism models of breakdown have, therefore, tended to support the traditional inferences. Crucially, however, connectionist models have greatly improved understanding of what modularity might look like in a neurocomputational system: a partial rather than an absolute property; a property that is the consequence of a developmental process where emergent specialization is driven by *structure-function correspondences* (the ability of certain parts of a computational structure to learn certain kinds of computation

better than other kinds); and a property that must now be complemented by concepts such as division of labor, degeneracy, interactivity, compensation, and redundancy (see Thomas & Karmiloff-Smith, 2002a). These insights have emerged even while advances in neuroimaging have tended to revise the overall notion of modularity, from an *a priori* theoretical principle of cognitive design to a data-driven way of describing patterns of activation across the brain during behavior (Thomas & Brady, 2021).

The most recent developments in cognitive neuropsychology have tended to reflect a growing trend in connectionist cognitive models as a whole: the inclusion of more constraints from neuroanatomy (Chen, Lambon Ralph, & Rogers, 2017). This produces so-called *connectivity-constrained* theories of cognition. For example, models of language have included dual pathways linking auditory areas for hearing a word to motor areas for producing the same word, reflecting the dorsal and ventral pathways observed in the brain (Ueno et al., 2011). This model is able to capture patterns of breakdown where adults can retain the ability to repeat words while losing the ability to comprehend them. Models of semantics have incorporated a hub-and-spoke architecture, where information from different sensory modalities is bound together in an amodal hub, based on the connectivity observed in the ventral anterior temporal lobe, the hub, with posterior fusiform gyrus (visual representations of objects), superior temporal gyrus (auditory representations of speech), and lateral parietal cortex (representations of object function and actions), the spokes (Chen et al., 2017). This model is able to capture various patterns of knowledge loss during semantic aphasia and semantic dementia as structure is lost from the anterior temporal lobe, as well as disorders stemming from the loss of control in retrieving semantic knowledge (Chen et al., 2017; Hoffman, McClelland, & Lambon Ralph, 2018). Lastly, the connectionist framework has been applied to the diagnosis of acquired disorders of language (Abel, Huber, & Dell, 2009) and therapeutic interventions (Abel, Willmes, & Huber, 2007), though the latter is comparatively under-developed to date (Thomas et al., 2019).

#### **2.4.4 The Origins of Individual Differences**

The fact that many connectionist models learn their cognitive abilities makes them a useful framework within which to study variations in trajectories of cognitive development, such as those associated with developmental disorders, intelligence, and giftedness. Connectionist models contain a number of constraints (architecture, activation dynamics, input and output representations, learning algorithm, training regime) that determine the efficiency and outcome of learning. Developmental outcomes may also be influenced by the quality of the learning experiences (the training set) to which the system is exposed. Manipulations to these constraints produce candidate explanations for impairments found in developmental disorders – for example, if a network has insufficient computational resources – or the impairments caused by exposure to

atypical environments such as in cases of deprivation, as well as the factors that underlie resilience and strong developmental outcomes.

In the 1980s and 1990s, many theories of developmental deficits employed the same explanatory framework as adult cognitive neuropsychology. There was a search for specific behavioral deficits or dissociations in children, which were then explained in terms of the failure of individual modules to develop. However, as Karmiloff-Smith (1998) pointed out, this meant that developmental deficits were actually being explained with reference to non-developmental, static, and sometimes adult models of normal cognitive structure. Karmiloff-Smith (1998, 2009) argued that the causes of developmental deficits of a genetic origin are likely to lie in changes to low-level neurocomputational properties that only exert their influence on cognition via an extended atypical developmental process (Elman et al., 1996; Mareschal et al., 2007). Connectionist models provided a way to explore the thesis that an understanding of the constraints on the developmental process is essential for generating explanations of developmental deficits. Models were applied to explaining a range of behavioral disorders including dyslexia, developmental language disorder and autism, as well as genetic disorders such as Williams syndrome and Down syndrome (Harm & Seidenberg, 1999; Joanisse & Seidenberg, 2003; Seidenberg, 2017; Thomas & Karmiloff-Smith, 2002b, 2003a; Thomas et al., 2016; Tovar, Westermann, & Torres, 2017).

If one can capture the development of the “average child,” and one can capture particular cases of atypical development, the stage is set to consider the origin of variations across the normal range. Some children develop more quickly than others; at a given age, a “bell-curve” or normal distribution of variation in ability is observed. The causes of such individual differences are often construed in terms of multiple interacting genetic and environmental factors. From the genetic side, the current view is that there are small contributions from many, perhaps thousands, of gene variants to individual differences in cognition, the so-called polygenic model (Knopik et al., 2016). From the environmental side, the most salient predictor of variation in cognitive outcomes is socio-economic status, although this metric is a proxy for potentially many underlying environmental influences (Hackman, Farah, & Meaney, 2010). To capture this range of variation in a formal model, however, requires simulations of whole populations, where individuals differ in their neurocomputational properties and in the quality of the learning environment to which they are exposed.

Connectionist models of cognitive development have been scaled to considering population-level characteristics in this manner, including applications to consider intelligence and giftedness (Thomas, 2016, 2018), as well as the interplay of genetic factors and of socio-economic status in influencing trajectories of development (Thomas, Forrester, & Ronald, 2013, 2016). These models have given mechanistic insight into how, for example, similar behavioral developmental disorders can arise from a *monogenic* cause – a large alteration of a single computational parameter produced by a genetic mutation – or from a

*polygenic* cause – the cumulative contribution of smaller differences in many computational parameters, perhaps lying on a continuum with variation in the normal range and produced by common genetic variants (Thomas & Knowland, 2014; Thomas et al., 2019).

Reflecting a move towards neuroanatomically constrained models discussed in the previous section, *multiscale* models of variation have sought to reconcile population-level data at multiple levels of description, including genes, brain structure, behavior, and environment (Thomas, Forrester, & Ronald, 2016). For example, to the extent that scientists are committed to viewing cognition as arising from the information processing properties of the brain, *genetic effects on cognition must correspond to influences on neurocomputational properties*; and some properties of connectionist networks, such as the number or strength of connections, can be seen as analogues to measures of brain structure, such as volumes of gray and white matter (Thomas, 2016). To give one recent example, Dündar-Coecke and Thomas (2019) sought to reconcile apparently paradoxical data from brain and behavior. Why are high IQs associated with having a bigger brain (as if more neural resources were better for cognition) but also associated with faster gray matter loss and cortical thinning during cognitive development (as if fewer neural resources were better for cognition)? The model suggested that the network size drives ability (so more is always better), but that a higher peak of network size during growth is then associated with faster connectivity loss as the brain optimizes processing through pruning unused resources (in the manner that higher mountain peaks have steeper sides).

Lastly, as with acquired disorders, implemented models of developmental deficits provide a foundation to explore interventions to ameliorate these deficits. While models of interventions are fewer than models of deficits, more attention has recently been paid to their implications. In these models, the success of behavioral interventions to remediate development deficits depends on the nature of the computational deficit, where it occurs in the model's architecture, the timing when the intervention is applied, and the content of the intervention items with respect to the training set (the latter corresponding to natural or educational experiences) (Thomas et al., 2019). Interventions that buttress developmental strengths rather than attempt to remediate weaknesses may also have more lasting benefits (Alireza, Fedor, & Thomas, 2017). These models may contribute to the (sometimes substantial) gap between theories of deficit and theories of treatment (see Moutoussis et al., 2017 for related work).

### 2.4.5 Deep Neural Networks for Cognitive Modeling

Deep neural networks have provided a step change in the performance of artificial intelligence systems for visual object recognition and natural language processing. Do they provide the basis for better cognitive models? As a case study, a number of researchers have explored whether the representations developed in the respective hidden unit layers of deep neural networks of visual object recognition accord to the types of representation found in the hierarchy of neural areas

in the ventral pathway of vision in inferior temporal cortex (e.g., Kriegeskorte, 2015; Yamins et al., 2014). Such a comparison is made possible by assessing the *representational similarity* between activity produced by a range of images of objects (faces, places, animals, tools, etc.), either in functional magnetic resonance imaging data of human participants or in the hidden unit activation levels of the trained neural network. The sequence of lower level features (edges), intermediate level features (contours), and high-level features (objects) is found both in neural areas and in network layers moving further from the input, suggesting similar computations are taking place. However, in other respects, these deep neural networks are not human-like: in the face of noise, their performance declines in nonhumanlike ways, suggesting over-fitting to the training data or the absence of crucial human-like architectural constraints; and at best, current models are capturing bottom-up, feedforward aspects of visual processing, not the top-down expectation-based influences enabled by bidirectional connectivity (Kriegeskorte, 2015; Storrs & Kriegeskorte, 2019).

Deep neural networks may be necessary to train more complex connectionist architectures suggested by the inclusion of neuroanatomical constraints. For example, Blakeman and Mareschal (2020) used a deep reinforcement learning architecture to model the interaction between neocortical, hippocampal, and striatal systems for learning the evaluation of actions. However, deep architectures do not provide better models solely by virtue of greater computational power. Indeed, the emergence of deep neural networks has resurrected some of the concerns expressed in early PDP days, that the lack of transparency in how trained networks operate limits their use for cognitive theory – if it is unknown how the model is working, how can the understanding of cognition be advanced? (See Seidenberg, 1993, for discussion.)

Some argue that deep neural networks are less readily extendible to higher level cognition, because unlike visual object recognition, it is unknown what cost function is being optimized (Aru & Vincente, 2018). For example, Aru and Vincente (2018) give the example of theory of mind/mindreading. The skills presumably being optimized (communication or deception) are themselves complex and hard to formulate. Higher cognitive functions may arise from the combination of many different neural processes that obey their own optimization cost functions. Others argue that deep networks indicate researchers in the field should ready themselves to deal with mechanisms that elude a concise mathematical description and an intuitive understanding (Kriegeskorte, 2015). The brain, after all, is complex. Yet others argue that understanding how big artificial neural networks work after they have learned will be similar to figuring out how the brain works but with several advantages: in the model, the following are known: exactly what each neuron computes, the learning algorithm they are using, and exactly how they are connected; the input can be controlled and the behavior of any set of neurons observed over an extended time period; and the system can be manipulated without any ethical concerns. Furthermore, these models may even be amendable to the methods used in cognitive psychology experiments (Ritter, Barrett, Santoro, & Botvinick, 2017).

### 2.4.6 Connectionism and Predictive Coding

Deep neural networks represent one instance of the reemergence of connectionism in the 2000s. Another can be identified in predictive processing, which has attracted considerable attention in certain areas of psychology, neuroscience, and philosophy. The idea of predictive coding was articulated in a paper on visual processing by Rao and Ballard (1999). Rao and Ballard proposed a model of visual processing in which feedback connections from a higher-order to a lower-order visual cortical area carry predictions of lower-level neural activities. This aspect of the predictive coding approach has similarities to the bidirectional, interlevel constraint satisfaction in McClelland & Rumelhart's (1981) Interactive Activation model of letter perception described in Section 2.3.1.

The broad idea of predictive processing is that a good internal model of the world will be one which can predict future sensory input. This will include the outcome of the organism's actions on the world on what will subsequently be perceived. And one way of improving the internal model is to compare its predictions against the actual sensory input and modify the model to reduce the disparity. This idea of minimizing temporal prediction error is already present in the SRN model of Elman (1990) described in Section 2.3.3, and is used widely in neural network models of learning and development.

However, predictive coding goes further in proposing that the signals propagated forward in the brain are prediction error signals; that is, only deviations from top-down expectations are passed between levels of representation within the sensory systems of the brain. Moreover, it proposes a role for precision weighting – a flexible calibration of how much noise is expected in bottom-up signals in a given context – in determining whether a disparity between top-down expectations and bottom-up input is sufficiently large to cause the internal model to update, so that it better predicts sensory input in the future. In the related idea of active inference, motor actions are no longer viewed as commands to move muscles but as descending predictions about proprioceptive sensory information (Friston, 2009).

The predictive coding approach has interesting applications to computational psychiatry, perception and action, although accounts of cognition formulated within this approach are not often used to create implemented models which capture details of human performance. While predictive coding shares features with some connectionist/PDP approaches, there are subtle differences whose empirical consequences remain to be worked out (see, e.g., Magnuson, Li, Luthra, You, & Steiner, 2019, for first steps in this direction).

## 2.5 Conclusion

This chapter has considered the foundation of connectionist modeling and its contribution to understanding of cognition. Connectionism was placed in the historical context of nineteenth-century associative theories of mental

processes and twentieth-century attempts to understand the computations carried out by networks of neurons, as well as the most recent innovations in deep learning. The key properties of connectionist networks were then reviewed, and particular emphasis placed on the use of learning to build the microstructure of these models. The core connectionist themes were: (1) that processing is simultaneously influenced by multiple sources of information at different levels of abstraction, operating via soft constraint satisfaction; (2) that representations are spread across multiple simple processing units operating in parallel; (3) that representations are graded, context-sensitive, and the emergent product of adaptive processes; (4) that computation is similarity-based and driven by the statistical structure of problem domains, but it can nevertheless produce rule-following behavior. The connectionist approach was illustrated via three foundational cognitive models, the Interactive Activation model of letter recognition (McClelland & Rumelhart, 1981), the past tense model (Rumelhart & McClelland, 1986), and simple recurrent networks for finding structure in time (Elman, 1990). Apart from its body of successful individual models, connectionist theory has had a widespread influence on cognitive theorizing, and this influence was illustrated by considering connectionist contributions to understanding of memory, cognitive development, acquired cognitive impairments, and cognitive variation. New emerging themes were identified, including connectionist models that incorporate neuroanatomical constraints, models that consider variation across populations reflecting the interaction of genetic and environmental influences, models that attempt to integrate data across levels of description, and models that make use of deep neural network architectures.

One could argue that since the first edition of this volume, a number of the theoretical constructs introduced by the connectionist approach have become so integrated into mainstream cognitive science, spurred by supporting evidence from neuroimaging, that they are no longer accompanied by the label “connectionist” – among them, notions like distributed representations shaped by task context; the role of prediction; and interactive processing (Mayor et al., 2014). Connectionism continues to challenge symbolic conceptions of thought, in areas such as language and mathematical cognition and in doing so, provides a more sympathetic framework for capturing developmental change. Recent directions have sought to integrate further constraints, such as from neuroanatomy and genetics. The future of connectionism, therefore, is likely to rely on its relationships with other fields within the cognitive sciences, and its ability to mediate between different levels of description in furnishing an understanding of the mechanistic basis of thought.

### Acknowledgments

This work was supported by grants from the Baily Thomas Charitable Fund, UK Medical Research Council [MR/R00322X/1]; and Fondation Jérôme Lejeune [2019b – 1901].

## References

- Abel, S., Huber, W., & Dell, G. S. (2009). Connectionist diagnosis of lexical disorders in aphasia. *Aphasiology*, *23*(11), 1353–1378.
- Abel, S., Willmes, K., & Huber, W. (2007). Model-oriented naming therapy: testing predictions of a connectionist model. *Aphasiology*, *21*(5), 411–447.
- Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science*, *9*, 147–169.
- Alireza, H., Fedor, A., & Thomas, M. S. C. (2017). Simulating behavioural interventions for developmental deficits: when improving strengths produces better outcomes than remediating weaknesses. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar, (Eds.), *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*, London, UK, July 26–29, 2017.
- Anderson, J., & Rosenfeld, E. (1988). *Neurocomputing: Foundations of Research*. Cambridge, MA: MIT Press.
- Anderson, J. A. (1977). Neural models with cognitive implications. In D. LaBerge & S. J. Samuels, (Eds.), *Basic Processes in Reading Perception and Comprehension*, pp. 27–90. Hillsdale, NJ: Erlbaum.
- Aru, J., & Vincente, R. (2018). What deep learning can tell us about higher cognitive functions like mindreading? *arXiv:1803.10470v2*
- Bechtel, W., & Abrahamsen, A. (1991). *Connectionism and the Mind*. Oxford: Blackwell.
- Berko, J. (1958). The child's learning of English morphology. *Word*, *14*, 150–177.
- Betti, A., & Gori, M. (2020). Backprop diffusion is biologically plausible. *arXiv:1912.04635v2*
- Blakeman, S., & Mareschal, D. (2020). A complementary learning systems approach to temporal difference learning. *Neural Networks*, *22*, 218–230. <https://doi.org/10.1016/j.neunet.2019.10.011>
- Botvinick, M. & Plaut, D. C. (2004). Doing without schema hierarchies: a recurrent connectionist approach to normal and impaired routine sequential action. *Psychological Review*, *111*, 395–429.
- Botvinick, M. M., & Cohen, J. D. (2014). The computational and neural basis of cognitive control: charted territory and new frontiers. *Cognitive Science*, *38*, 1249–1285. <https://doi.org/10.1111/cogs.12126>
- Brown, T. B., Mann, B., Ryder, N., et al. (2020) Language models are few-shot learners. arXiv preprint. *arXiv:2005.14165*.
- Burton, A. M., Bruce, V., & Johnston, R. A. (1990). Understanding face recognition with an interactive activation model. *British Journal of Psychology*, *81*, 361–380.
- Bybee, J., & McClelland, J. L. (2005). Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition. *The Linguistic Review*, *22*(2–4), 381–410.
- Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, *113*(2), 234–272. <https://doi.org/10.1037/0033-295X.113.2.234>
- Chen, P. L., Lambon Ralph, M., & Rogers, T. T. (2017). A unified model of human semantic knowledge and its disorders. *Nature Human Behaviour*, *1*, 0039. <https://doi.org/10.1038/s41562-016-0039>
- Christiansen, M. H. & Chater, N. (2001). *Connectionist Psycholinguistics*. Westport, CT: Ablex.



- Cleeremans, A., & Dienes, Z. (2008). Computational models of implicit learning. In R. Sun (Ed.), *The Cambridge Handbook of Computational Psychology* (pp. 396–421). Cambridge: Cambridge University Press. <https://doi.org/10.1017/cbo9780511816772.018>
- Cobb, M. (2020). *The Idea of the Brain*. London: Profile Books.
- Cohen, G., Johnstone, R. A., & Plunkett, K. (2000). *Exploring Cognition: Damaged Brains and Neural Networks*. Hove: Psychology Press.
- Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: a parallel distributed processing account of the Stroop effect. *Psychological Review*, *97*, 332–361.
- Crick, F. (1989). The recent excitement about neural networks. *Nature*, *337*, 129–132. <https://doi.org/10.1038/337129a0>
- Davelaar, E. J., & Usher, M. (2002). An activation-based theory of immediate item memory. In J. A. Bullinaria & W. Lowe (Eds.), *Proceedings of the Seventh Neural Computation and Psychology Workshop: Connectionist Models of Cognition and Perception*. Singapore: World Scientific.
- Davies, M. (2005). Cognitive science. In F. Jackson & M. Smith (Eds.), *The Oxford Handbook of Contemporary Philosophy*. Oxford: Oxford University Press.
- Devlin, J., Gonnerman, L., Andersen, E., & Seidenberg, M.S. (1997). Category specific semantic deficits in focal and widespread brain damage: a computational account. *Journal of Cognitive Neuroscience*, *10*, 77–94.
- Dündar-Coecke, S., & Thomas, M. S. C. (2019). Modeling socioeconomic effects on the development of brain and behavior. In A. K. Goel, C. M. Seifert, & C. Freksa (Eds.), *Proceedings of the 41<sup>st</sup> Annual Conference of the Cognitive Science Society* (pp. 1676–1682). Montreal: Cognitive Science Society.
- 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*, 195–224.
- Elman, J. L. (1993). Learning and development in neural networks: the importance of starting small. *Cognition*, *48*, 71–99.
- Elman, J. L. (2005). Connectionist models of cognitive development: where next? *Trends in Cognitive Sciences*, *9*, 111–117.
- Elman, J. L. & McRae, K. (2019). A model of event knowledge. *Psychological Review*, *126* (2), 252–291. <https://doi.org/10.1037/rev0000133>
- 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.
- Ervin, S. M. (1964). Imitation and structural change in children's language. In E. H. Lenneberg (Ed.), *New Directions in the Study of Language*. Cambridge, MA: MIT Press.
- Fahlman, S., & Lebiere, C. (1990). The cascade correlation learning architecture. In D. Touretzky (Ed.), *Advances in Neural Information Processing 2* (pp. 524–532). Los Altos, CA: Morgan Kaufman.
- Feldman, J. A. (1981). A connectionist model of visual memory. In G. E. Hinton & J. A. Anderson (Eds.), *Parallel Models of Associative Memory* (pp. 49–81). Hillsdale, NJ: Erlbaum.

- Fitz, H., & Chang, F. (2017). Meaningful questions: the acquisition of auxiliary inversion in a connectionist model of sentence production. *Cognition*, *166*, 225–250. <https://doi.org/10.1016/j.cognition.2017.05.008>
- Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: a critical analysis. *Cognition*, *78*, 3–71.
- French, R. M., Ans, B., & Rousset, S. (2001). Pseudopatterns and dual-network memory models: advantages and shortcomings. In R. French & J. Sougné (Eds.), *Connectionist Models of Learning, Development and Evolution* (pp. 13–22). London: Springer.
- Freud, S. (1895). Project for a scientific psychology. In J. Strachey (Ed.), *The Standard Edition of the Complete Psychological Works of Sigmund Freud*. London: The Hogarth Press and the Institute of Psycho-Analysis.
- Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends in Cognitive Sciences*, *13*(7), 293–301. <https://doi.org/10.1016/j.tics.2009.04.005>
- Friston, K., & Kiebel, S. (2009). Predictive coding under the free-energy principle. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *364*(1521), 1211–1221. <https://doi.org/10.1098/rstb.2008.0300>
- Goebel, R., & Indefrey, P. (2000). A recurrent network with short-term memory capacity learning the German –s plural. In P. Broeder & J. Murre (Eds.), *Models of Language Acquisition: Inductive and Deductive Approaches* (pp. 177–200). Oxford: Oxford University Press.
- Gordon, P. (2004). Numerical cognition without words: evidence from Amazonia. *Science*, *306*(5695), 496–499.
- Grainger, J., Midgley, K., & Holcomb, P. J. (2010). Re-thinking the bilingual interactive-activation model from a developmental perspective (BIA-d). In M. Kail & M. Hickmann (Eds.), *Language Acquisition Across Linguistic and Cognitive Systems* (pp. 267–283). Amsterdam: John Benjamins Publishing Company.
- Green, D. C. (1998). Are connectionist models theories of cognition? *Psychology*, *9*(4).
- Grossberg, S. (1976a). Adaptive pattern classification and universal recoding I: parallel development and coding of neural feature detectors. *Biological Cybernetics*, *23*, 121–134.
- Grossberg, S. (1976b). Adaptive pattern classification and universal recoding II: feedback, expectation, olfaction, and illusions. *Biological Cybernetics*, *23*, 187–202.
- Haarmann, H., & Usher, M. (2001). Maintenance of semantic information in capacity limited item short-term memory. *Psychonomic Bulletin & Review*, *8*, 568–578.
- Hackman, D. A., Farah, M. J., & Meaney, M. J. (2010). Socioeconomic status and the brain. *Nature Reviews Neuroscience*, *11*, 651–659.
- Hahnloser, R., Sarpeshkar, R., Mahowald, M., et al. (2000). Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit. *Nature*, *405*, 947–951. <https://doi.org/10.1038/35016072>
- Harm, M. W. & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: insights from connectionist models. *Psychological Review*, *106* (3), 491–528.
- Hebb, D. O. (1949). *The Organization of Behavior: A Neuropsychological Approach*. New York, NY: John Wiley & Sons.
- Hinton, G. E. (1989). Deterministic Boltzmann learning performs steepest descent in weight-space. *Neural Computation*, *1*, 143–150.

- Hinton, G. E., & Anderson J. A. (1981). *Parallel Models of Associative Memory*. Hillsdale, NJ: Erlbaum.
- 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, NY: American Institute of Physics.
- Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313* (5786), 504–507.
- Hinton, G. E., & Sejnowski, T. (1986). Learning and relearning in Boltzmann machines. In D. Rumelhart & J. McClelland (Eds.), *Parallel Distributed Processing* (vol. 1, pp. 282–317). Cambridge, MA: MIT Press.
- Hinton, G. E., & Sejnowski, T. J. (1983). Optimal perceptual inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Washington, DC.
- Hochreiter, S. (1991). *Untersuchungen zu dynamischen neuronalen Netzen*. Diploma thesis, Institut f. Informatik, Technische Univ. Munich.
- Hochreiter, S., Bengio, Y., Frasconi, P., & Schmidhuber, J. (2001). Gradient flow in recurrent nets: the difficulty of learning long-term dependencies. In S. C. Kremer & J. F. Kolen (Eds.), *A Field Guide to Dynamical Recurrent Neural Networks*. Piscataway, NJ: IEEE Press.
- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Hoeffner, J. H., & McClelland, J. L. (1993). Can a perceptual processing deficit explain the impairment of inflectional morphology in developmental dysphasia? A computational investigation. In E. V. Clark (Ed.), *Proceedings of the 25th Child Language Research Forum* (pp. 38–49). Stanford, CA: Center for the Study of Language and Information.
- Hoffman, P., McClelland, J., & Lambon Ralph, M. (2018). Concepts, control and context: a connectionist account of normal and disordered semantic cognition. *Psychological Review*, *125*(3), 293–328. <https://doi.org/10.1037/rev0000094>
- Hofstadter, D. (2018). The Shallowness of Google Translate. *The Atlantic*. Available from: [www.theatlantic.com/technology/archive/2018/01/the-shallowness-of-google-translate/551570/](http://www.theatlantic.com/technology/archive/2018/01/the-shallowness-of-google-translate/551570/) [last accessed August 9, 2022].
- Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Science USA*, *79*, 2554–2558.
- Houghton, G. (2005). *Connectionist Models in Cognitive Psychology*. Hove: Psychology Press.
- James, W. (1890). *Principles of Psychology*. New York, NY: Holt.
- Joanisse, M. F. & McClelland, J. L. (2015). Connectionist perspectives on language learning, representation, and processing. *WIREs Cognitive Science* (online). <https://doi.org/10.1002/wcs.1340>
- Joanisse, M. F. & Seidenberg, M. S. (1999). Impairments in verb morphology following brain injury: a connectionist model. *Proceedings of the National Academy of Science*, *96*, 7592–7597.
- Joanisse, M. F. & Seidenberg, M. S. (2003). Phonology and syntax in specific language impairment: evidence from a connectionist model. *Brain and Language*, *86*, 40–56.

- Jordan, M. I. (1986). Attractor dynamics and parallelism in a connectionist sequential machine. In *Proceedings of the Eighth Annual Conference of Cognitive Science Society* (pp. 531–546). Hillsdale, NJ: Erlbaum.
- Karaminis, T. N., & Thomas, M. S. C. (2010). A cross-linguistic model of the acquisition of inflectional morphology in English and Modern Greek. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, August 11–14, 2010. Portland, Oregon, USA.
- Karaminis, T. N., & Thomas, M. S. C. (2014). The multiple inflection generator: a generalized connectionist model for cross-linguistic morphological development. *DNL Tech report 2014* (online). [http://193.61.4.246/dnl/wp-content/uploads/2020/04/KT\\_TheMultipleInflectionGenerator2014.pdf](http://193.61.4.246/dnl/wp-content/uploads/2020/04/KT_TheMultipleInflectionGenerator2014.pdf) [last accessed August 9, 2022].
- Karmiloff-Smith, A. (1998). Development itself is the key to understanding developmental disorders. *Trends in Cognitive Sciences*, 2, 389–398.
- Karmiloff-Smith, A. (2009). Nativism versus neuroconstructivism: rethinking the study of developmental disorders. *Developmental Psychology*, 45(1), 56–63.
- Kirov, C. & Cotterell, R. (2018). Recurrent neural networks in linguistic theory: revisiting Pinker and Prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics*, 6, 651–665. [https://doi.org/10.1162/tacl\\_a\\_00247](https://doi.org/10.1162/tacl_a_00247)
- Knopik V. S., Neiderhiser J. M., DeFries J.C., & Plomin R. (2016). *Behavioral genetics* (7th ed). New York, NY: Worth Publishers.
- Kohonen, T. (1984). *Self-Organization and Associative Memory*. Berlin: Springer-Verlag.
- Kollias, P. & McClelland, J. L. (2013). Context, cortex, and associations: a connectionist developmental approach to verbal analogies. *Frontiers in Psychology*, 4, 857. <https://doi.org/10.3389/fpsyg.2013.00857>
- Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, 1, 417–446.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Proceedings of the 25th International Conference on Neural Information Processing Systems*, 1, 1097–1105.
- Kuczaj, S. A. (1977). The acquisition of regular and irregular past tense forms. *Journal of Verbal Learning and Verbal Behavior*, 16, 589–600.
- Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40, e253.
- Lashley, K. S. (1929). *Brain Mechanisms and Intelligence: A Quantitative Study of Injuries to the Brain*. New York, NY: Dover Publications, Inc.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521 (7553), 436.
- Lillicrap, T., Cownden, D., Tweed, D., & Akerman, C. J. (2016). Random synaptic feedback weights support error backpropagation for deep learning. *Nature Communications*, 7, 13276. <https://doi.org/10.1038/ncomms13276>
- Lillicrap, T. P., Santoro, A., Marris, L., Akerman, C. J., & Hinton, G. E. (2020). Backpropagation and the brain. *Nature Reviews Neuroscience*, 21, 335–346. <https://doi.org/10.1038/s41583-020-0277-3>

- MacDonald, M. C., & Christiansen, M. H. (2002). Reassessing working memory: a comment on Just & Carpenter (1992) and Waters & Caplan (1996). *Psychological Review*, *109*, 35–54.
- MacKay, D. J. (1992). A practical Bayesian framework for backpropagation networks. *Neural Computation*, *4*, 448–472.
- Magnuson, J. S., Li, M., Luthra, S., You, H., & Steiner, R. (2019). Does predictive processing imply predictive coding in models of spoken word recognition? In *Proceedings of the 41th Annual Meeting of the Cognitive Science Society* (pp. 735–740). Cognitive Science Society.
- Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, *117*(48), 30046–30054.
- Marcus, G. F. (2001). *The Algebraic Mind: Integrating Connectionism and Cognitive Science*. Cambridge, MA: MIT Press.
- Marcus, G., Pinker, S., Ullman, M., Hollander, J., Rosen, T., & Xu, F. (1992). Overregularisation in language acquisition. *Monographs of the Society for Research in Child Development*, *57* (228), 1–178.
- Mareschal, D., & Thomas M. S. C. (2007). Computational modeling in developmental psychology. *IEEE Transactions on Evolutionary Computation (Special Issue on Autonomous Mental Development)*, *11*, 137–150.
- Mareschal, D., Johnson, M., Sirios, S., Spratling, M., Thomas, M. S. C., & Westermann, G. (2007). *Neuroconstructivism: How the Brain Constructs Cognition*. Oxford: Oxford University Press.
- Marr, D. (1982). *Vision*. San Francisco, CA: W. H. Freeman.
- Marr, D., & Poggio, T. (1976). Cooperative computation of stereo disparity. *Science*, *194*, 283–287.
- Mayor, J., Gomez, P., Chang, F., & Lupyan, G. (2014). Connectionism coming of age: legacy and future challenges. *Frontiers In Psychology*, *5*, 187. <https://doi.org/10.3389/fpsyg.2014.00187>
- McClelland, J. L. (1981). Retrieving general and specific information from stored knowledge of specifics. In *Proceedings of the Third Annual Meeting of the Cognitive Science Society* (pp. 170–172). Hillsdale, NJ: Lawrence Erlbaum Associates.
- McClelland, J. L. (1989). Parallel distributed processing: implications for cognition and development. In M. G. M. Morris (Ed.), *Parallel Distributed Processing, Implications for Psychology and Neurobiology* (pp. 8–45). Oxford: Clarendon Press.
- McClelland, J. L. (2013). Integrating probabilistic models of perception and interactive neural networks: a historical and tutorial review. *Frontiers in Psychology*, *4*, 503. [www.frontiersin.org/articles/10.3389/fpsyg.2013.00503/full](http://www.frontiersin.org/articles/10.3389/fpsyg.2013.00503/full)
- McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1–86.
- McClelland, J. L., Hill, F., Rudolph, M., Baldridge, J., & Schuetze, H. (2020). Placing language in an integrated understanding system: next steps toward human-level performance in neural language models. *Proceedings of the National Academy of Sciences*, *117*(42), 25966–25974. <https://doi.org/10.1073/pnas.1910416117>
- 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., Plaut, D. C., Gotts, S. J., & Maia, T. V. (2003). Developing a domain-general framework for cognition: what is the best approach? Commentary on a target article by Anderson and Lebiere. *Behavioral and Brain Sciences*, 22, 611–614.
- 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(5), 375–405.
- McClelland, J. L., Rumelhart, D. E. & the PDP Research Group (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 2: Psychological and Biological Models*. Cambridge, MA: MIT Press.
- McCulloch, W. S., & Pitts, W. (1943). A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5, 115–133. Reprinted in Anderson & Rosenfield (1988).
- McLeod, P., Plunkett, K., & Rolls, E. T. (1998). *Introduction to Connectionist Modelling of Cognitive Processes*. Oxford: Oxford University Press
- Meynert, T. (1884). *Psychiatry: A Clinical Treatise on Diseases of the Forebrain. Part I. The Anatomy, Physiology and Chemistry of the Brain*. Trans. B. Sachs. New York, NY: G. P. Putnam's Sons.
- Minsky, M., & Papert, S. (1969). *Perceptrons: An Introduction to Computational Geometry*. Cambridge, MA: MIT Press.
- Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, 76, 165–178.
- Morton, J. B., & Munakata, Y. (2002). Active versus latent representations: a neural network model of perseveration, dissociation, and decalage in childhood. *Developmental Psychobiology*, 40, 255–265.
- Moutoussis, M., Shahar, N., Hauser, T. U., & Dolan, R. J. (2017). Computation in psychotherapy, or how computational psychiatry can aid learning-based psychological therapies. *Computational Psychiatry*, 2, 50–73. [https://doi.org/10.1162/cpsy\\_a\\_00014](https://doi.org/10.1162/cpsy_a_00014)
- Movellan, J. R., & McClelland, J. L. (1993). Learning continuous probability distributions with symmetric diffusion networks. *Cognitive Science*, 17, 463–496.
- Munakata, Y. (1998). Infant perseveration and implications for object permanence theories: a PDP model of the AB task. *Developmental Science*, 1, 161–184.
- Munakata, Y. & McClelland, J. L. (2003). Connectionist models of development. *Developmental Science*, 6, 413–429.
- Newell, A. (1980). Physical symbol systems. *Cognitive Science*, 4(2), 135–183.
- Novikoff, A. (1962). *Proceedings of the Symposium on the Mathematical Theory of Automata*, 12, 615–622. New York, NY, USA, Polytechnic Institute of Brooklyn.
- O'Reilly, R. C. (1996). Biologically plausible error-driven learning using local activation differences: the generalized recirculation algorithm. *Neural Computation*, 8, 895–938.
- O'Reilly, R. C. (1998). Six principles for biologically based computational models of cortical cognition. *Trends in Cognitive Sciences*, 2, 455–462.
- O'Reilly, R. C., Bhattacharyya, R., Howard, M. D., & Ketz, N. (2014). Complementary learning systems. *Cognitive Science*, 38, 1229–1248. <https://doi.org/10.1111/j.1551-6709.2011.01214.x>

- O'Reilly, R. C., Braver, T. S., & Cohen, J. D. (1999). A biologically based computational model of working memory. In A. Miyake, & P. Shah (Eds.), *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*. New York, NY: Cambridge University Press.
- O'Reilly, R. C., & Munakata, Y. (2000). *Computational Explorations in Cognitive Neuroscience: Understanding the Mind by Simulating the Brain*. Cambridge, MA: MIT Press.
- Pater, J. (2019). Generative linguistics and neural networks at 60: foundation, friction, and fusion. *Language, 95*(1). Epub February 20, 2019. <https://doi.org/10.1353/lan.2019.0005>
- Piazza, M., Pica, P., Izard, V., Spelke, E. S., & Dehaene, S. (2013). Education enhances the acuity of the nonverbal approximate number system. *Psychological Science, 24*(6), 1037–1043. <https://doi.org/10.1177/0956797612464057>.
- Pinker, S. (1984). *Language Learnability and Language Development*. Cambridge, MA: Harvard University Press.
- Pinker, S. (1999). *Words and Rules*. London: Weidenfeld & Nicolson.
- 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., & Kello, C. T. (1999). The emergence of phonology from the interplay of speech comprehension and production: a distributed connectionist approach. In B. MacWhinney (Ed.), *The Emergence of Language* (pp. 381–415). Mahwah, NJ: Erlbaum.
- Plaut, D. C. & McClelland, J. L. (1993). Generalization with componential attractors: word and nonword reading in an attractor network. In *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* (pp. 824–829). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. E. (1996). Understanding normal and impaired word reading: computational principles in quasi-regular domains. *Psychological Review, 103*, 56–115.
- Plunkett, K., & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perceptron: implications for child language acquisition. *Cognition, 38*, 1–60.
- Plunkett, K., & Marchman, V. (1993). From rote learning to system building: acquiring verb morphology in children and connectionist nets. *Cognition, 48*, 21–69.
- Plunkett, K., & Marchman, V. (1996). Learning from a connectionist model of the English past tense. *Cognition, 61*, 299–308.
- Plunkett, K., & Nakisa, R. (1997). A connectionist model of the Arabic plural system. *Language and Cognitive Processes, 12*, 807–836.
- Rao, R., & Ballard, D. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience, 2*, 79–87. <https://doi.org/10.1038/4580>
- Rashevsky, N. (1935). Outline of a physico-mathematical theory of the brain. *Journal of General Psychology, 13*, 82–112.
- Reicher, G. M. (1969). Perceptual recognition as a function of meaningfulness of stimulus material. *Journal of Experimental Psychology, 81*, 274–280.
- Ritter, S., Barrett, D. G. T., Santoro, A., & Botvinick, M. M. (2017). Cognitive psychology for deep neural networks: a shape bias case study. *arXiv:1706.08606v2*

- Rohde, D. L. T. & Plaut, D. C. (1999). Language acquisition in the absence of explicit negative evidence: how important is starting small? *Cognition*, 72, 67–109.
- Rosenblatt, F. (1958). The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65, 386–408.
- Rosenblatt, F. (1962). *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Washington, DC: Spartan Books
- Rumelhart, D. E., & McClelland, J. L. (1982). An interactive activation model of context effects in letter perception. Part 2: The contextual enhancement effect and some tests and extensions of the model. *Psychological Review*, 89, 60–94.
- Rumelhart, D. E., & McClelland, J. L. (1985). Levels indeed! *Journal of Experimental Psychology General*, 114(2), 193–197.
- 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, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Volume 1: Foundations* (pp. 318–362). Cambridge, MA: MIT Press.
- Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. In D. E. Rumelhart, J. L. McClelland, & the PDP Research Group, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations* (pp. 45–76). Cambridge, MA: MIT Press.
- Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tense of English verbs. In J. L. McClelland, D. E. Rumelhart, & the PDP Research Group (Eds.). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 2: Psychological and Biological Models* (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, Vol. 1: Foundations*. Cambridge, MA: MIT Press.
- Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). *Schemata and sequential thought processes in PDP models*. In J. L. McClelland, D. E. Rumelhart, & the PDP Research Group, *Explorations in the Microstructure of Cognition Volume 2: Psychological and Biological Models* (p. 7–57). Cambridge, MA: MIT Press.
- Sabatiel, S., McClelland, J. L., & Solstad, T. (2020). A computational model of learning to count in a multimodal, interactive environment. *Proceedings of the 42nd Annual Conference of the Cognitive Science Society*. Austin, TX: Cognitive Science Society.
- Saffran, J. R., & Kirkham, N. Z. (2018). Infant statistical learning. *Annual Review of Psychology*, 69, 181–203. <https://doi.org/10.1146/annurev-psych-122216-011805>
- Saffran, J. R., Newport, E. L., & Aslin, R. N. (1996). Word segmentation: the role of distributional cues. *Journal of Memory and Language*, 35, 606–621.
- Scellier, B., & Bengio, Y. (2019). Equivalence of equilibrium propagation and recurrent backpropagation. *Neural Computation*, 31(2), 312–329. [https://doi.org/10.1162/neco\\_a\\_01160](https://doi.org/10.1162/neco_a_01160)
- Schmidhuber, J. (2015). Deep learning in neural networks: an overview. *Neural Networks*, 61, 85–117. <https://doi.org/10.1016/j.neunet.2014.09.003>



- Seidenberg, M. S. (1993). Connectionist models and cognitive theory. *Psychological Science, 4*(4), 228–235.
- Seidenberg, M. S. (2017). *Language at the Speed of Sight*. New York, NY: Basic Books.
- Selfridge, O. G. (1959). Pandemonium: a paradigm for learning. In *Symposium on the Mechanization of Thought Processes* (pp. 511–529). London: HMSO.
- Shallice, T. (1988). *From Neuropsychology to Mental Structure*. Cambridge: Cambridge University Press.
- Shultz, T. R. (2003). *Computational Developmental Psychology*. Cambridge, MA: MIT Press.
- Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences, 11*, 1–74.
- Spencer, H. (1872). *Principles of Psychology* (3rd ed.). London: Longman, Brown, Green, & Longmans.
- Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research, 15*, 1929–1958.
- Stoianov, I., & Zorzi, M. (2012). Emergence of a ‘visual number sense’ in hierarchical generative models. *Nature Neuroscience, 15*(2), 194–196.
- Storrs, K. R., & Kriegeskorte, N. (2019). Deep learning for cognitive neuroscience. *arXiv:1903.01458v1*
- Sutton, R. S., & Barto, A. G. (1981). Toward a modern theory of adaptive networks: expectation and prediction. *Psychological Review, 88*(2), 135–170.
- Testolin, A., Zou, W. Y., & McClelland, J. L. (2020). Numerosity discrimination in deep neural networks: initial competence, developmental refinement and experience statistics. *Developmental Science, 2020*, e12940.
- Thomas, M. S. C. (2016). Do more intelligent brains retain heightened plasticity for longer in development? A computational investigation. *Developmental Cognitive Neuroscience, 19*, 258–269. <https://doi.org/10.1016/j.dcn.2016.04.002>
- Thomas, M. S. C. (2018). A neurocomputational model of developmental trajectories of gifted children under a polygenic model: when are gifted children held back by poor environments? *Intelligence, 69*, 200–212.
- Thomas, M. S. C., & Brady, D. (2021). Quo vadis modularity in the 2020s? In M. S. C. Thomas, D. Mareschal, & V. C. P. Knowland (Eds). *Taking Development Seriously: A Festschrift for Annette Karmiloff-Smith*. London: Routledge Psychology.
- Thomas, M. S. C., Davis, R., Karmiloff-Smith, A., Knowland, V. C. P., & Charman, T. (2016). The over-pruning hypothesis of autism. *Developmental Science, 9*(2), 284–305. <https://doi.org/10.1111/desc.12303>
- Thomas, M. S. C., Fedor, A., Davis, R., Yang, J., Alireza, H., Charman, T., Masterson, J., & Best, W. (2019). Computational modelling of interventions for developmental disorders. *Psychological Review, 26*(5), 693–726. <https://doi.org/10.1037/rev0000151>
- Thomas, M. S. C., Forrester, N. A., & Richardson, F. M. (2006). What is modularity good for? In *Proceedings of The 28th Annual Conference of the Cognitive Science Society* (pp. 2240–2245), July 26–29, Vancouver, BC, Canada.
- Thomas, M. S. C., Forrester, N. A., & Ronald, A. (2013). Modeling socioeconomic status effects on language development. *Developmental Psychology, 49*(12), 2325–2343. <https://doi.org/10.1037/a0032301>

- Thomas, M. S. C., Forrester, N. A., & Ronald, A. (2016). Multi-scale modeling of gene-behavior associations in an artificial neural network model of cognitive development. *Cognitive Science*, *40*(1), 51–99. <https://doi.org/10.1111/cogs.12230>
- Thomas, M. S. C., & Karmiloff-Smith, A. (2002a). Are developmental disorders like cases of adult brain damage? Implications from connectionist modelling. *Behavioral and Brain Sciences*, *25*(6), 727–788.
- Thomas, M. S. C., & Karmiloff-Smith, A. (2002b). Modelling typical and atypical cognitive development. In U. Goswami (Ed.), *Handbook of Childhood Development* (pp. 575–599). Oxford: Blackwell.
- Thomas, M. S. C., & Karmiloff-Smith, A. (2003a). Modeling language acquisition in atypical phenotypes. *Psychological Review*, *110*(4), 647–682.
- Thomas, M. S. C., & Karmiloff-Smith, A. (2003b). Connectionist models of development, developmental disorders and individual differences. In R. J. Sternberg, J. Lautrey, & T. Lubart (Eds.), *Models of Intelligence: International Perspectives*, (pp. 133–150). Washington, DC: American Psychological Association.
- Thomas, M. S. C., & Knowland, V. C. P. (2014). Modelling mechanisms of persisting and resolving delay in language development. *Journal of Speech, Language, and Hearing Research*, *57*(2), 467–483. [https://doi.org/10.1044/2013\\_JSLHR-L-12-0254](https://doi.org/10.1044/2013_JSLHR-L-12-0254)
- Thomas, M. S. C., & Van Heuven, W. (2005). Computational models of bilingual comprehension. In J. F. Kroll & A. M. B. De Groot (Eds.), *Handbook of Bilingualism: Psycholinguistic Approaches* (pp. 202–225). Oxford: Oxford University Press.
- Touretzky, D. S., & Hinton, G. E. (1988). A distributed connectionist production system. *Cognitive Science*, *12*, 423–466.
- Tovar, A., Westermann, G., & Torres, A. (2017). From altered LTP/LTD to atypical learning: a computational model of Down syndrome. *Cognition*, *171*, 15–24. <https://doi.org/10.1016/j.cognition.2017.10.021>
- 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. <https://doi.org/10.1016/j.neuron.2011.09.013>
- Usher, M., & McClelland, J. L. (2001). On the time course of perceptual choice: the leaky competing accumulator model. *Psychological Review*, *108*, 550–592.
- van Gelder, T. (1991). Classical questions, radical answers: connectionism and the structure of mental representations. In T. Horgan & J. Tienson (Eds.), *Connectionism and the Philosophy of Mind*. Dordrecht: Kluwer Academic Publishers.
- Verguts, T., & Fias, W. (2004). Representation of number in animals and humans: a neural model. *Journal of Cognitive Neuroscience*, *16*(9), 1493–1504. <https://doi.org/10.1162/0898929042568497>
- Westermann, G., Mareschal, D., Johnson, M. H., Sirois, S., Spratling, M. W., & Thomas, M. S. C. (2007). Neuroconstructivism. *Developmental Science*, *10*, 75–83.
- Westermann, G., Thomas, M. S. C., & Karmiloff-Smith, A. (2010). Neuroconstructivism. In U. Goswami (Ed.), *Blackwell Handbook of Child Development* (2nd ed.), (pp. 723–748). Oxford: Blackwell.
- Williams, R. J., & Zipser, D. (1995). Gradient-based learning algorithms for recurrent networks and their computational complexity. In Y. Chauvin & D. E.

- Rumelhart (Eds.), *Back-propagation: Theory, Architectures and Applications*. Hillsdale, NJ: Erlbaum.
- Woollams, A. M. (2014). Connectionist neuropsychology: uncovering ultimate causes of acquired dyslexia. *Philosophical Transactions of the Royal Society B*, 369 (1634), <https://doi.org/10.1098/rstb.2012.0398>
- Wu, Y., Schuster, M., Chen, Z., et al. (2016). Google's neural machine translation system: bridging the gap between human and machine translation. Available from: <https://arxiv.org/abs/1609.08144> [last accessed August 9, 2022].
- Xie, X., & Seung, H. S. (2003). Equivalence of backpropagation and contrastive Hebbian learning in a layered network. *Neural Computation*, 15, 441–454.
- Xu, F., & Pinker, S. (1995). Weird past tense forms. *Journal of Child Language*, 22, 531–556.
- Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23), 8619–8624.