

Psychological Function in Computational Models of Neural Networks

Randall C. O'Reilly
Department of Psychology
University of Colorado at Boulder
Campus Box 345
Boulder, CO 80309-0345
(303) 492-0054 (2967 fax)
oreilly@psych.colorado.edu

Yuko Munakata
Department of Psychology
University of Denver
2155 S. Race St.
Denver, CO 80208
(303) 871-4151 (4747 fax)
munakata@du.edu

1st February 2001

Draft chapter for the *Handbook of Psychology, Vol 3, Biological Psychology*
Michela Gallagher & Randy Nelson, Eds. New York: Wiley

Introduction

The overarching goal of cognitive neuroscience is to understand how the brain gives rise to thought. Toward this goal, researchers employ various methods to measure neural variables while people and other animals think. A complementary method, computer models of neural networks, allows unparalleled levels of control and supports the further understanding of the relation between brain and mind. Using these models, one can simulate a network of interacting neurons, and measure cognitive function in these networks at the same time. Furthermore, many variables in these networks can be manipulated, so that their effects on cognitive processes can be observed.

In this chapter, we provide an up-to-date review of some of the core principles and prominent applications of computational models in cognitive neuroscience, based on our recent textbook on this topic (O'Reilly & Munakata, 2000). We begin with a summary of some of the basic questions confronting computational modelers in cognitive neuroscience. We then discuss provisional answers to these questions, showing how they apply to a range of empirical data. Throughout, and in closing, we discuss challenges to neural network models. We will see how some network models can have possibly problematic properties, often driven by constraints from biology or cognition, but the models can nonetheless help to advance the field of cognitive neuroscience.

Basic Mechanistic Questions Facing Computational Models

As soon as one is faced with the task of constructing a neural network from scratch, several important questions immediately arise:

- How do the neurons talk to each other? We know a lot about the answers to this question from neuroscience, but there are some more specific questions that models raise, including:
 - What kind of information do spikes contain — is it just the rate of spiking, or are more detailed signals being conveyed in the timing of individual spikes?
 - How are spike inputs from other neurons integrated together within the neuron — does each input just add into an overall sum, or are different inputs treated differently?
 - Are there basic network-level patterns of interaction between neurons that apply across a wide range of cognitive functions?
- How do networks learn from experience? Networks with even a few tens of neurons can exhibit complex and varied patterns of behavior depending on how the neurons are interconnected. Brain sized networks with billions of neurons are thus almost incomprehensibly complex. The brain requires a powerful way of setting all of these patterns of interconnectivity to achieve useful behaviors in the face of all the other random possibilities.
- How does the myriad of complex perceptual inputs get organized into a coherent internal representation of the environment?
- How are memories of previous experience stored, organized, and retrieved?
- How does higher level cognition arise from networks of neurons?

There are many other such questions that one could ask, but we'll focus on relatively brief treatments of these due to space limitations (see O'Reilly & Munakata, 2000 for a comprehensive treatment of a wide range of cognitive phenomena). In the process of addressing each of these questions, we develop a set of basic mechanisms for computational cognitive neuroscience modeling.

The Neural Activation Function

The first two questions we raised above (“What kind of information do spikes contain?” and “How are spike inputs from other neurons integrated together within the neuron?”) can be addressed by developing what is commonly called an *activation function* for a simulated neuron. This activation function provides a mathematical formalism (i.e., an *algorithm*) for describing how neurons talk to each other. The “currency” of this neural communication is referred to as “activation” — neurons communicate by activating each other. Fortunately, neuroscience has developed a relatively advanced understanding the basic operation of a neuron (reviewed in Chapter ?? of this volume??). We summarize the key facts here.

A neuron receives input from other neurons through *synapses* located on branching processes called *dendrites*. These inputs are somehow integrated together as an electrical voltage (referred to as the *membrane potential*) in the *cell body*. If this voltage exceeds a certain value (called the *threshold*), then the neuron will trigger a *spike* of electrical activity down its *axon*, which is another branching process that terminates on the synapses of other neurons. This input to other neurons thus continues the propagation of information through the brain.

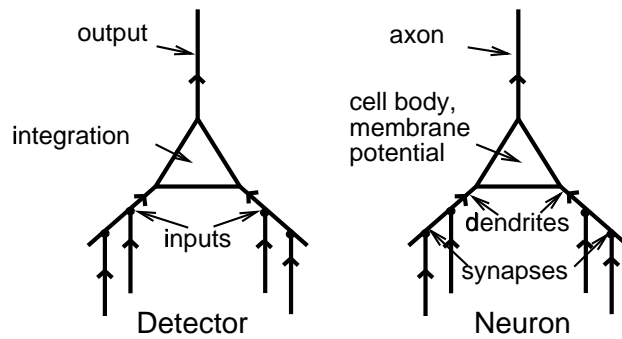


Figure 1: The detector model can be used to understand the function of corresponding neural components.

All of these properties are central to the simulated neurons in neural network models. One can frame these properties in terms of a *detector* — this detector model provides a clear functional role for the biological properties of the neuron (O’Reilly & Munakata, 2000). Specifically, we can think of a neuron as detecting the existence of some set of conditions, and responding with a signal that communicates the extent to which those conditions have been met. Think of a smoke detector, which is constantly sampling the air looking for conditions that indicate the presence of a fire. In the brain, there are neurons in the early stages of the visual system that are constantly sampling the visual input looking for conditions that indicate the presence of very simple visual features such as bars of light at a given position and orientation in the visual scene. Higher up in the visual system, there are neurons that detect different sets of objects. One can interpret virtually any neural activation in the brain in the general language of detecting some kind of pattern of activity in the inputs.

The match between the detector model and the biological features of the neuron is shown in figure 1. The dendrites provide detector inputs, which get integrated into an overall signal in the cell body that the neuron can use to determine if what it is detecting is actually there. The firing threshold acts just like a smoke detector threshold — there needs to be sufficient evidence of a fire before an alarm should be triggered, and similarly the neuron must accumulate a sufficient voltage before it can send information to other neurons.

A critical aspect of simulated neurons (also called *units*) is that they have adjustable parameters called *weights* that determine how much influence the different inputs have on them. In the neuron, these weights correspond to the strength of the individual synapses connecting neurons — some synapses produce more electrical voltage input than others. Thus, some inputs “weigh” more heavily into the detection decision than do others.

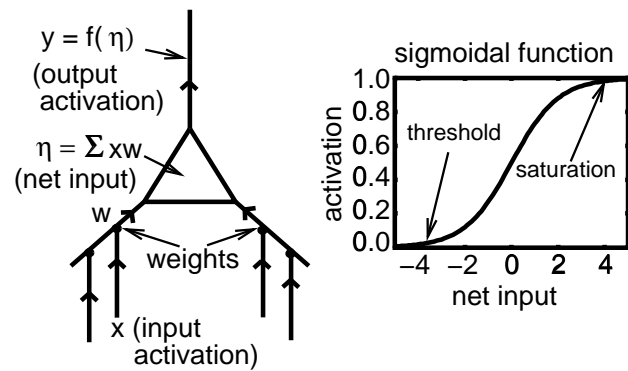


Figure 2: Basic equations for simple models of neurons. Input activation from other neurons x is multiplied times the weight value w , and summed together to get the net input ($\eta = \sum xw$). The output activation y is then computed as a function $f(\eta)$ of the net input. In the simplest kind *linear* model, there is no output function, $y = \eta$. A more complex model has a *threshold*, below which activation does not occur: $y = \eta$ if $\eta > \Theta$; 0 otherwise. A saturation point can also be imposed, so that activation y has a maximal limit. The *sigmoidal* function $f(\eta) = \frac{1}{1+e^{-\eta}}$ produces a graded, continuous-valued function that has both threshold-like and saturating properties.

These weights provide the critical parameters for specifying what the neuron detects — essentially, neurons can detect *patterns of activity* over their inputs, with those input patterns that best fit the pattern of weights producing the largest detection response. A perhaps more tangible example of this kind of operation would be obtained by looking through a piece of paper with a shape cut out of it (e.g., an L shape). If you look through this “mask” at a black computer screen with different shapes being displayed on it in a bright color, you’ll see the most light if the displayed shape exactly matches the cutout shape. Similarly, a neuron gets the most “activation” if its inputs match its pattern of weights.

There is a progression of increasingly more complex ways of simulating the basic activation function of a neuron (figure 2). In the simplest kind of neural model, a unit simply adds up the weighted inputs from other neurons (called the *net input* to a unit), and sends the results directly on to other neurons. This is a *linear unit*. It turns out that linear units, like linear equations, are not very powerful in a computational sense — the effects of layers and layers of linear units can all be summarized by just one such layer. The next most complex kind of unit is a *threshold linear unit*, which is just like a linear unit except it has a threshold (much like a real neuron), so that if its net input is below the threshold, it does not send any activation to other units. Because neurons can only fire so fast, it also makes sense to think of there being a *saturation* point in the activation function, above which

the activation cannot go. The widely-used *sigmoidal unit* (figure 2) provides a more graded function that exhibits both threshold-like and saturating properties.

These simplified models of the neural activation function are at the heart of a controversy surrounding the nature of the “neural code”. Specifically, in mapping these models onto the brain, the simulated neural activations are real-valued numbers, so we have to assume that they represent something like the *average firing rate* of neurons, not the individual spikes themselves. Thus, a major question is: Does the firing rate capture the essence of the information that real neurons communicate to each other, or does the precise timing of the spikes contain more information that would be lost with a rate code? The debate over the nature of the information encoded in the neural spike train has inspired a number of empirical studies across a range of different animals, with some demonstrations that detailed spike timing matters in some parts of the brain (e.g., Reike, Warland, van Steveninck, & Bialek, 1996). However, recordings in the cortex of primates, which is the most relevant for understanding human cognition, have been largely consistent with the rate code simplification in neural networks (Tovee, Rolls, Treves, & Bellis, 1993).

Moving beyond these highly simplified model neurons, one can incorporate equations that simulate the electrical behavior of biological neurons. It has long been known that the electrical properties of neurons can be understood in terms of the familiar concepts of resistors and capacitors, the so-called *equivalent circuit* model of the neuron. A key issue that arises here is the extent to which a neuron behaves like a single coherent electrical system, or whether one needs to keep track of electrical potentials at many different locations along the neuron. At the simplest extreme is a *point neuron*, which has the entire geometry of the neuron reduced to a single point in space, such that only one electrical potential needs to be computed. These biologically-based units are not much more complex than the units just described, and yet they do a better job of capturing the behavior of real neurons — for this reason, we used the point neuron model in O'Reilly and Munakata (2000). Much more complex neural models have also been explored using many different electrical *compartments* (e.g., Koch & Segev, 1998).

The debate over how many compartments to use in a neural model centers on the issue of how real neurons integrate all of their inputs. If neurons are effectively electrically unitary (as in a point neuron), then they essentially add their inputs together, just as in the simplified models. If, however, different parts of the neuron have very different electrical potentials, then inputs coming into these different parts can have very different ef-

fects, and therefore a simple additive integration would underestimate the real complexity of neurons (e.g., Shepherd & Brayton, 1987). Although it is by no means the final word on these issues, a recent analysis (Jaffe & Carnevale, 1999) supports the idea that the point neuron model captures much of the integration properties of real cortical neurons. Specifically, they found that as long as there was not a long primary dendrite (as in most cortical pyramidal neurons), the impact of inputs at various places on the dendritic tree was roughly the same when measured at the soma.

One important biological constraint that the point neuron captures but the simpler sigmoidal-style units do not is that excitation and inhibition are separated in the brain. Thus, a given neuron will either send an excitatory or an inhibitory output to other neurons. Excitatory outputs result in the excitation of other neurons, while inhibitory outputs counteract this excitation and make the receiving neuron less likely to become activated. In the simpler sigmoidal unit, weights are typically allowed to be either positive or negative, which violates this biological constraint.

In summary, one could argue that the point neuron model captures the essential properties of individual neural computations. Although this model undoubtedly commits errors of omission by not capturing many details of real neurons, in some cases the functional relevance of such details may be approximated by the point neuron model (e.g., the functional relevance of a synapse may boil down to its efficacy, which can be approximated by a weight parameter, without capturing all of the biological details of actual synapses). In other cases, such details may not be all that functionally important. At the least, such a simple model does not commit obvious errors of commission; that is, it does not attribute any computational powers to the model neurons that are unlikely to be true of real neurons.

Network Interactions

In this section, we move beyond the individual neuron and consider the properties of networks of interconnected neurons, with the next of our overarching questions in mind: “Are there basic network-level patterns of interaction between neurons that apply across a wide range of cognitive functions?” By identifying basic network-level interactions, we can develop a vocabulary of mechanistically-grounded concepts for understanding how neural networks behave.

Before identifying these network properties, we need to know what the cortical network looks like. Figure 3 shows how the 6-layered structure of the cortex varies in different areas of the brain. Based on this informa-

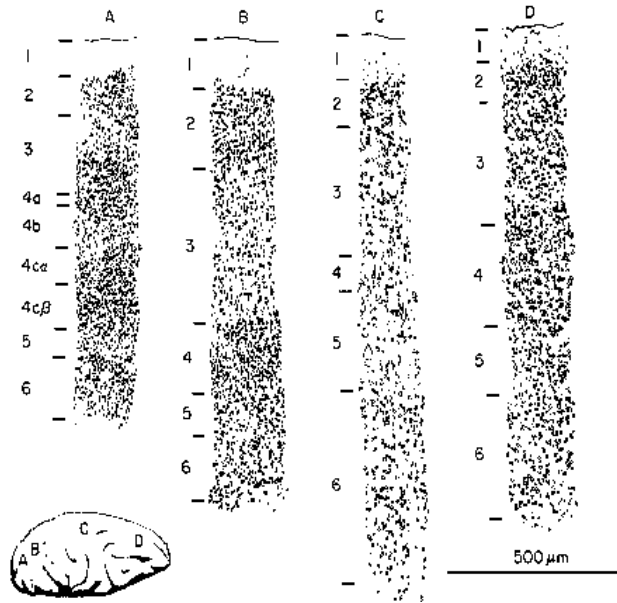


Figure 3: Different laminar structure of cortex from different areas. A) shows specialization for the input layer 4 in the primary visual input area. B) shows emphasis on hidden layers 2/3 in a hidden area (extrastriate cortex) higher up in the visual processing stream. C) shows emphasis on output layers 5/6 in a motor output area. D) shows a relatively even blend of layers in a prefrontal area. Reproduced from Shepherd (1990).

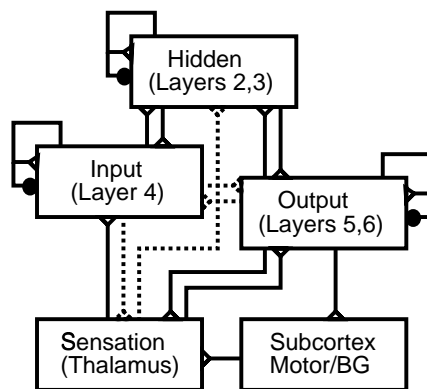


Figure 4: A simple, three-layer interpretation of cortical structure that is consistent with general connectivity patterns and provides a useful starting point for modeling. Direct excitatory connectivity is shown by the open triangular connections. Inhibitory interneurons are indicated by the filled circular connections; these operate within each cortical layer and receive the same types of excitatory connections as the excitatory neurons do. Dotted lines indicate connections that may exist but are not consistent with the feedforward flow of information from input to hidden to output (with feedback projections along the same pathways). Limited data make it difficult to determine how prevalent and important these connections are.

tion and anatomical studies of connectivity, one can summarize the roles of the 6-layered cortical structure in terms of the following as three *functional* layers (figure 4; O'Reilly & Munakata, 2000):

- The *input* layer corresponds to cortical layer 4, which usually receives the sensory input by way of a subcortical brain area called the *thalamus*, which receives information from the retina and other sense organs.
- The *output* layer corresponds to the *deep* cortical layers 5 and 6, which send motor commands and other outputs to a wide range of subcortical areas, including the *basal ganglia*.
- The *hidden* layer (so called because it is not directly “visible” via either the inputs or outputs) corresponds to the *superficial* (upper) cortical layers 2 and 3 (cortical layer 1 is largely just axons). These layers receive inputs locally from the other cortical layers and also from other more distant superficial cortical layers. The superficial layers usually project outputs back to these same more distant cortical areas, and to the deep (output) layers locally.

Thus, at a very coarse level, sensory information flows into the input layer, gets “processed” in some way by the hidden layer, and the results of this processing give rise to motor outputs in the output layer. As figure 3 shows, this trajectory of information flow is actually spread across many different cortical areas, each of which has the same 6-layered structure. However, input areas tend to have a more pronounced layer 4, output areas have more pronounced layers 5 and 6, and hidden (association) areas have more pronounced layers 2 and 3. Figure 5 summarizes these layer-level specializations in different brain areas, as they contribute to the “big loop” of information flow through the brain.

Thus, based on the structure of the cortex, it is reasonable for neural network models to be composed of the three functional layers (input, one or more hidden layers, and an output layer). The biology also tells us that the primary neuron type (the *pyramidal neuron*) that connects between layers and between brain areas is excitatory, while there are a number of inhibitory *interneurons* that have more local patterns of connectivity. The excitatory connectivity is generally *bidirectional*, meaning that excitation flows from input to hidden to output, and at the same time backwards from output to hidden to input.

Using these features of the biology, we can frame the basic network-level interactions in terms of three different kinds of neural connectivity between functional layers of neurons (O'Reilly & Munakata, 2000):

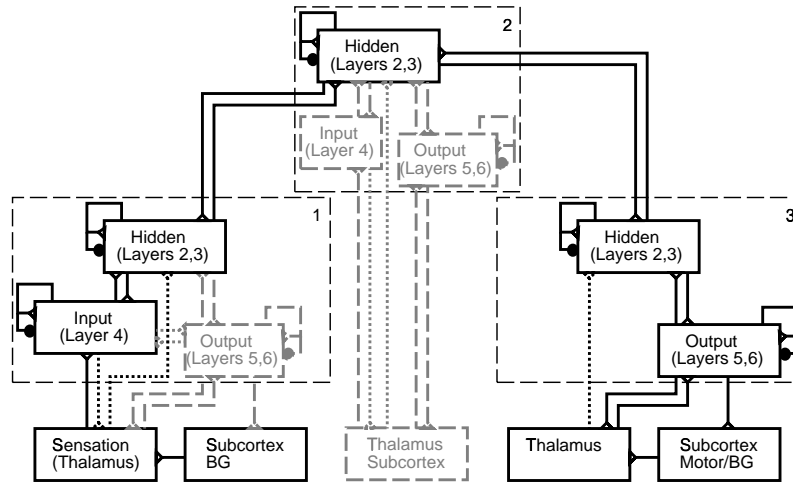


Figure 5: Larger scale version of cortical circuitry, showing the three different types of cortical areas: **1**) is an *input area*, which has a well-developed layer 4 receiving sensory input from the thalamus, but not producing motor output directly. **2**) is a *hidden area* (often called a higher-level association area), which receives input from input areas (or other hidden areas) and sends outputs to output areas (or other hidden areas). It has reduced input and output layers, and communicates primarily via layer 2–3 connectivity. **3**) is an *output area* (motor control area), projecting to subcortical brain areas that drive the motor system, which has no real layer 4 (and a larger layer 5). Dashed lines indicate layers and connections that are reduced in importance for a given area, and dotted lines again represent connections which may exist but are not consistent with the input-hidden-output model of feedforward information flow. In all cases, reciprocal backwards projections complement the feedforward projections, enabling bidirectional, interactive processing.

- Unidirectional (feedforward) excitation: when one layer of neurons sends activation to another layer, the patterns of activation in the sending layer get *transformed*, by *emphasizing* some distinctions while *collapsing* across others. Each neural detector contributes to the transformation by its own selectivities in what it detects and what it ignores. These transformations are the basic information processing operations in the brain, creating ever more abstract and powerful *categories* of information over subsequent layers of processing.
- Bidirectional (feedback) excitation: when two layers each send activation to the other (also called *interactivity* or *recurrence*), the layers can amplify each other's activations, giving rise to a variety of phenomena including *pattern completion*, *resonance*, *mutual support*, and *attractor dynamics*. This also allows for the feedforward-style transformations to proceed in both directions.
- Inhibitory competition: feedforward communication of inhibition across layers and feedback inhibition within a layer combine to produce a competition among neurons within a layer — only those neurons receiving the strongest amount of excitation will remain active in the face of this competition.

The overall picture is this: Sensory information gets processed by a cascade of hidden layers, where each hidden layer extracts some particular kind of information, while discarding others. For example, it is well known that different parts of the visual system extract color, motion, object identity, and object location information. Extracting this information requires carefully-tuned patterns of weights that emphasize the relevant information while ignoring the irrelevant. For example, where the visual image of an object appears on retina is irrelevant for identifying what the object is, but differences in the shape of the image are critical. Conversely, shape information is irrelevant for locating that object in space, but retinal position (along with eye, head, and body positions) is relevant. Essentially, the brain extracts ever more abstract “executive summaries” of the environment, and then uses this information to formulate plans of action, which are then executed through projections to the output layers.

Superimposed upon this input-output flow of information are the complex dynamics that arise from bidirectional connectivity and inhibition. Bidirectional connectivity enables the brain to “fill in” missing input information using “higher level” knowledge. For example, a hidden area that encodes words at an abstract level can fill in the missing pieces of a garbled phone message by relying on knowledge of who is speaking, and what they are likely to be saying. In combination with inhibitory

competition, bidirectional connectivity also ensures that all the different brain areas focus on the same thing at the same time — neurons processing information relevant to other brain areas will be reinforced through the bidirectional excitation, and this extra excitation will lead to greater inhibition on other neurons that aren't relevant to other brain areas, thereby shutting them off. At a cognitive level, we can think of this interplay of bidirectional excitation and inhibition as giving rise to *attention* — we'll explore this idea in a subsequent section.

Learning Mechanisms

In many psychological theories, learning assumes a relatively tangential role — people assume that the brain systems they hypothesize arise through some complex interaction between genetic structures and experience, but the details of exactly how the systems came to be are often too difficult to confront while trying to provide a theory about the mature system. In contrast, learning plays an essential role in most neural network models, because generally the best way to get a neural network to do something useful is to have it learn based on its experiences (i.e., to change its weights according to a learning mechanism, based on patterns of activations that get presented to the input layer of the network). To see why, you must first appreciate that a network's behavior is determined by the patterns of weights over its units. Even relatively small networks have three layers (input, hidden, output), of say 20 units each, with full connectivity between layers, meaning that there are a minimum of 800 connection weights (400 for input-hidden connections and another 400 for hidden-output connections, with more if you include bidirectional connectivity back from the output to the hidden, and any kind of inhibitory connectivity that might be used). It is virtually impossible to set such a large number of weights by hand, so learning takes center stage. For the same reason, neural networks have proven to be a useful tool for exploring a range of developmental issues (Elman, Bates, Johnson, Karmiloff-Smith, Parisi, & Plunkett, 1996; Munakata & Stedron, in press).

A considerable amount of research has been conducted on neural mechanisms of synaptic modification, which is the biological equivalent of changing the weights between neural units, as would be required by a learning mechanism. The overall finding from these studies is that connection strengths between neurons (weights) can be made to either go up or down depending on the relationship between the activations of the sending and receiving neurons. When the weight goes up, it is referred to as *long term potentiation* (LTP), and when it goes down it is called *long term depression* (LTD).

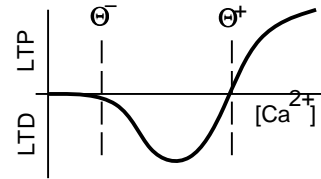


Figure 6: Relationship between LTP and LTD, where a moderate amount of increased intracellular calcium ion concentration leads to LTD, but at a larger amount leads to LTP.

One well-supported account of what causes LTP versus LTD is based on the concentration of calcium ions in the receiving neuron's dendrite in the vicinity of the synapse (Lisman, 1989, 1994; Bear & Malenka, 1994). Calcium ions can enter the dendrite through various means, but the dominant contributor in cortical neurons is probably the opening of NMDA channels located at excitatory synapses. Interestingly, NMDA channels open (to allow calcium ions to enter) only if two things happen: (a) the sending neuron fires and releases neurotransmitter that binds to the NMDA channel, and (b) the receiving neuron's electrical potential increases above a critical level. Thus, calcium ion concentration tends to reflect the extent to which both sending and receiving neurons are active. The calcium ion model further stipulates that LTP (weight increases) will occur when calcium ion concentration is very high, whereas LTD (weight decreases) occur if calcium ion concentration is elevated, but not so high (figure 6). If there is no change in calcium ion concentration, no weight change occurs. Putting this calcium mechanism together with the requirements for opening the NMDA channel, it is clear that weight increases should occur when both the sending and receiving neurons are strongly active together. Conditions where LTD will occur are less clear.

The biological mechanisms just described are remarkably consistent with Hebb's (1949) postulate that groups of neurons that are active together should increase the strength of their interconnectivity — the brain should learn about things that go together. In other words, under the *Hebbian* learning principle, which appears to be supported by biological synaptic modification mechanisms, the brain encodes *correlated* events. These correlations are meaningful because correlation is a good (though imperfect) clue for causation, and co-occurring items can be more efficiently represented together within a common representational structure (e.g., the concept "college dorm room" evokes a whole slew of co-occurring items like pizza boxes, posters, boom boxes, etc.).

Many computational models of Hebbian learning have been developed (e.g., Kohonen, 1984; Grossberg, 1976; Oja, 1982; Bienenstock, Cooper, & Munro, 1982;

Rumelhart & Zipser, 1986; Carpenter & Grossberg, 1987; Linsker, 1988; Miller, Keller, & Stryker, 1989), with applications to a range of different cognitive and neural phenomena. Despite the successes of these models, Hebbian learning has some significant limitations. Specifically, Hebbian learning cannot learn arbitrary input-output transformations (McClelland & Rumelhart, 1988; O'Reilly & Munakata, 2000). To see why this is an important limitation, we can refer back to the input-hidden-output network structure discussed in the previous section. In general, the organism's learning task can be construed as learning a set of hidden representations based on sensory inputs that produce useful output patterns (behavioral responses). Importantly, the relationship between sensory inputs and motor outputs can be highly complex and essentially arbitrary.

The limitations of Hebbian learning are most evident when compared with the other major form of neural network learning mechanism, *error driven learning*. In error-driven learning, the network's weights are adjusted according to the differences between the output pattern the network actually produced and the output pattern it should have produced (i.e., the error). So, if the network executes a pulling motion when it gets a "push" command, it can adjust the connections to specifically correct this error. Error-driven learning mechanisms have been around for a long time in one form or another (Widrow & Hoff, 1960), and have been applied to a wide range of animal learning phenomena (e.g., Rescorla & Wagner, 1972). However, these earlier versions were limited to learning connections between an input and output layer only — they could not handle the training of intermediate hidden layer representations. This limitation was seized upon by Minsky and Papert (1969) in their devastating critique showing that these neural networks were very limited in the kinds of input-output mappings they could learn, which had the effect of significantly curtailing research in this field. However, the extension of error-driven learning mechanisms to networks with (multiple) hidden layers via the *error-backpropagation* learning procedure, and the concomitant demonstration that these networks could learn virtually any input-output mapping, revived interest some 15 years later (Rumelhart, Hinton, & Williams, 1986; the idea had also been developed several times before Bryson & Ho, 1969; Werbos, 1974; Parker, 1985).

Error-driven mechanisms can learn many input-output mapping problems that Hebbian learning simply fails to learn (O'Reilly & Munakata, 2000). The reason is clear — Hebbian learning is designed to encode correlations, not to learn arbitrary input-output mappings. However, instead of arguing for the exclusive superiority of one learning mechanism over the other, one can obtain complementary benefits by using both kinds of

learning mechanisms (Hebbian and error-driven). This combination of both types of learning, together with an inhibitory competition mechanism, is the defining characteristic of the *Leabra* framework (O'Reilly, 1996b, 1998; O'Reilly & Munakata, 2000). In short, error-driven learning provides the ability to learn arbitrary input-output mappings, while Hebbian learning provides a useful tendency to encode correlated information. Furthermore, Hebbian learning acts locally at each neuron, and is therefore a relatively fast and reliable form of learning, whereas error-driven learning depends on distant error-signals that can become weak and unreliable as they propagate through multiple hidden layers.

One potential problem with the *Leabra* framework and all other network models that rely upon error-driven learning is a possible error of commission with respect to the known neurobiology. Indeed, much has been made in the literature about the biological implausibility of the error-backpropagation learning mechanism, which appears to require a type of signal that has never been measured in neurons to propagate in the reverse direction of most neural signals (e.g., Crick, 1989; Zipser & Andersen, 1988). Furthermore, it has not been clear where the necessary "desired outputs" for generating error signals could plausibly come from. However, it has recently been shown that bidirectional activation propagation (as discussed in the previous section) can be used to perform essentially the same error-driven learning as backpropagation (O'Reilly, 1996a), using any of a number of readily available teaching signals. The resulting algorithm generalizes the recirculation algorithm of Hinton & McClelland (Hinton & McClelland, 1988), and is thus called *GeneRec*. *GeneRec* provides the error-driven component of the *Leabra* algorithm.

The basic idea behind *GeneRec* is that instead of propagating an error signal, which is a difference between two terms, one can propagate the two terms separately as activation signals, and then take their difference locally at each unit. This works by having two phases of activations for computing the two terms. In the *expectation* phase, the bidirectionally-connected network processes an input activation pattern into a state that reflects the expected consequences or correlates of that input pattern. Then, in the *outcome* phase, the network experiences actual consequences or correlates. The difference between outcome and expectation is the error signal, and the bidirectional connectivity propagates this error signal throughout the network via local activation signals.

The *GeneRec* analysis also showed that Boltzmann machine learning and its deterministic versions (Ackley, Hinton, & Sejnowski, 1985; Hinton, 1989; Peterson & Anderson, 1987; Movellan, 1990) can be seen as variants of this more biologically plausible version

of the backpropagation algorithm. This means that all of the existing approaches to error-driven learning using activation-based signals converge on essentially the same basic mechanism, making it more plausible that this is the way the brain does error driven learning. Furthermore, the form of synaptic modification necessary to implement this algorithm is consistent with (though not directly validated by) the calcium-ion based synaptic modification mechanism described earlier. Finally, there are many sources in the natural environment for the necessary outcome phase signals in the form of actual environmental outcomes that can be compared with internal expectations to provide error signals (McClelland, 1994; O'Reilly, 1996a). Thus, one does not need to have an explicit “teacher” to perform error-driven learning.

To summarize, learning mechanisms are at once the most important and most controversial aspects of neural network models. In this discussion, we have seen that Hebbian learning mechanisms make close contact with biological mechanisms, whereas error-driven mechanisms have been motivated largely from top-down constraints from cognition — they are the only known mechanisms capable of learning the kinds of things that we know people can learn. The two kinds of mechanisms may be combined in a biologically plausible and powerful way.

Perceptual Processing and Attention

Having presented some of the most central ideas behind the basic mechanisms used in neural network models, we now turn to applications of these mechanisms for understanding cognitive phenomena. These same mechanisms have been applied to a wide variety of phenomena; we focus here on perception, attention, memory, and higher level cognition. The first question we address was stated in the introduction: “How does the myriad of complex perceptual inputs get organized into a coherent internal representation of the environment?”

We describe two different ways that neural network models have provided insight into this question. The first is by addressing the *representational* problem — what kinds of representations provide an efficient, computationally useful encoding of the perceptual world for a neural network, and do these representations look anything like those actually found in the brain? We will see that the interaction between Hebbian learning mechanisms and inhibitory competition can produce visual representations very much like those found in the brain. The second is by addressing the *attentional* problem — given that there is a huge overload of perceptual information impinging upon us at every moment (e.g., as you try to read this chapter), how does our brain focus on and select

out the most relevant information (hopefully this chapter!) for further processing? We will see that the interaction between inhibitory competition and bidirectional activation flow can produce emergent attentional dynamics that simulate the behavior of both intact and brain lesioned people on a visual attention task.

The Structure of Representations in Primary Visual Cortex

One way of understanding what representations in primary visual cortex (V1) should look like from a computational perspective is to simply present a range of visual images to a model network and allow its learning mechanisms to develop representations that encode these images. This is indeed what a number of modelers have done, using natural visual scenes that were preprocessed in a manner consistent with the contrast-enhancement properties of the retina (e.g., Olshausen & Field, 1996; Bell & Sejnowski, 1997; van Hateren & van der Schaaff, 1997; O'Reilly & Munakata, 2000). The Olshausen and Field (1996) model demonstrated that *sparse* representations (with relatively few active neurons) provide a useful basis for encoding real-world (visual) environments, but this model was not based on known biological principles. Subsequent work replicated the same general results using more biologically-based principles of Hebbian learning and sparseness constraints in the form of inhibitory competition between neurons (O'Reilly & Munakata, 2000). Furthermore, lateral excitatory connections within this network produced a *topographic organization* of representations, where neighboring units had similar representations.

Figure 7 shows the results from the O'Reilly and Munakata (2000) model of 14x14 hidden units (representing V1 neurons) receiving inputs from a 12x12 simulated “retina.” This figure shows that the simulated neurons have developed *oriented edge detectors*; the neurons are maximally activated by visual inputs that have transitions between dark and light regions separated by edges at various angles. We can understand why the network develops these receptive fields in terms of the proclivity of Hebbian learning to encode correlational structure. Natural objects tend to have piecewise linear edges, so that strong correlations exist among pixels of light along these edges. However, Hebbian learning alone is not enough to produce these receptive field patterns. As emphasized by Olshausen and Field (1996), a constraint of only having a relatively few units active at any time (implemented by inhibitory competition in our model) is also important. This constraint is appropriate because only a relatively small number of oriented edges are present in any given image. Furthermore, in the process of learning, inhibition ensures that units compete

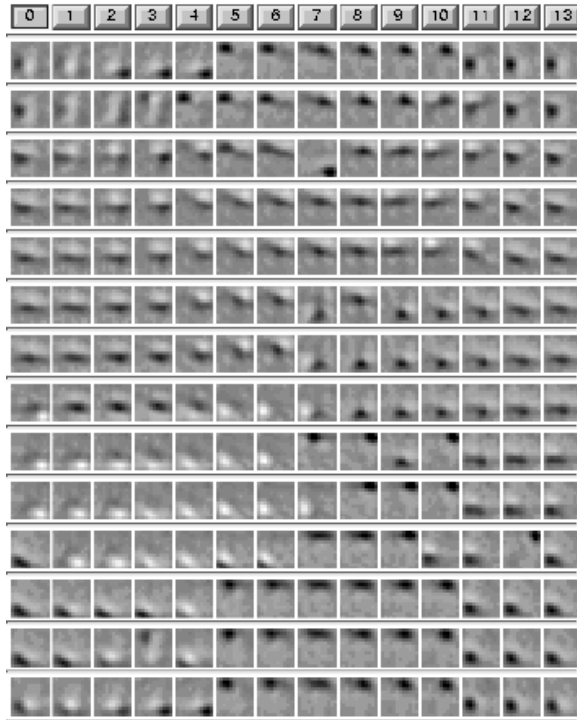


Figure 7: The receptive fields of model V1 neurons (from O'Reilly & Munakata, 2000). The broader 14x14 grid contains individual unit receptive fields, within which there is a smaller 12x12 grid representing weights from a simulated “retina”. Lighter shades indicate areas of on-center response, and darker shades indicate areas of off-center response to retinal inputs.

and specialize to represent different aspects of the input. At an intuitive level, this learning process is analogous to the effects of competition and natural selection in biological evolution (e.g., Edelman, 1987). Thus, each unit carves out a different “niche” in the space of all possible reliable correlations in the input images — these niches are oriented edge detectors.

This analysis shows that we can understand the general principles of why computational models develop their representations, and why these are appropriate for a given domain of input patterns. However, do these principles help us understand how the brain works? They can if the representations developed by the model look like those in the brain. It turns out that they do — V1 neurons have long been known to encode oriented edges of light (Hubel & Wiesel, 1962; Marr, 1982). Furthermore, one can find systematic variations in orientation, size, position, and *polarity* (i.e., going from light-to-dark or dark-to-light, or dark-light-dark and light-dark-light) in both the simulated and real V1 receptive fields. In the brain, the different types of edge detectors (together with other neurons that appear to encode visual surface properties) are packed into the two-dimensional

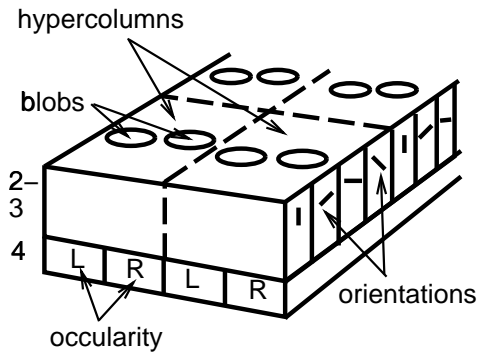


Figure 8: Structure of a cortical hypercolumn, that represents a full range of orientations (in layers 2–3), ocular dominance columns (in layer 4, one for each eye), and surface features (in the blobs). Each such hypercolumn is focused within one region of retinal space, and neighboring hypercolumns represent neighboring regions.

sheet of the visual cortex according to a *topographic* organization. The large-scale organization is a *retinotopic map* that preserves the topography of the retinal image in the cortical sheet. At the smaller scale are *hypercolumns* (figure 8) containing smoothly varying progressions of oriented edge detectors, among other things (Livingstone & Hubel, 1988). The topography shown in figure 7 is consistent with this within-hypercolumn structure. The hypercolumn also contains *ocular dominance columns*, in which V1 neurons respond preferentially to input from one eye or the other (see Miller et al., 1989 for a Hebbian-based model). For reviews of the many computational models of various of these V1 structures, see Swindale (1996) and Erwin, Obermayer, and Schulten (1995).

To summarize, computational models incorporating the basic mechanisms of Hebbian learning and inhibitory competition can help us understand *why* V1 has the representations it does.

Spatial Attention and the Effects of Parietal Lobe Damage

The dynamics of activation flow through the network are as important as the weight patterns of the neurons in the network. One of the most widely-studied manifestations of these dynamics is attention to different regions of visual space. Spatial attention has classically been operationalized according to the Posner spatial cuing task (Posner, Walker, Friedrich, & Rafal, 1984, figure 9). When attention is drawn or *cued* to one region of space, participants are then faster to detect a target in that region (a validly cued trial) than a target elsewhere (an invalidly cued trial). Patients with damage to the parietal lobe have particular difficulty with invalidly cued trials.



Figure 9: The Posner spatial attention task. The cue is a brightening or highlighting of one of the boxes that focuses attention to that region of space. Reaction times to detect the target are faster when this cue is valid (the target appears in that same region) than when it is invalid (the target appears elsewhere).

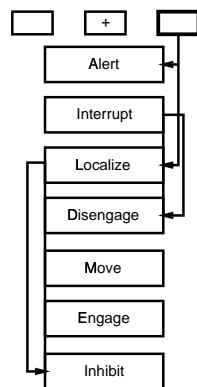


Figure 10: The information processing model for attentional processing according to Posner and colleagues.

The traditional account of the spatial attention data involves a sequence of modular processes that have been associated with different brain areas (Posner et al., 1984; figure 10). Specifically, the parietal brain damage data is accounted for in terms of a disengage module associated with the parietal lobe (Posner et al., 1984). This module typically allows one to disengage from an attended location to attend elsewhere. This process of disengaging takes time, leading to the slower detection of targets in unattended locations. Further, the disengage module is impaired with parietal damage, leading patients to have difficulty disengaging from attention drawn to one side of space.

Biologically-based computational models, based on reinforcing excitatory connections and competitive inhibitory connections, provide an alternative explanation for these phenomena (Cohen, Romero, Farah, & Servan-Schreiber, 1994; O'Reilly & Munakata, 2000). In this framework, the facilitory effects of drawing attention to one region of space result from bidirectional excitatory connections between spatial and other representations of that region — this excitatory support makes it easier to process information in that region because neurons are already receiving supporting excitation. The slowing

that comes on the invalid trials results from inhibitory competition between different spatial regions — to activate a different spatial location requires inhibiting the previously-active region. Under this model, damage to the parietal lobe simply impairs the ability of the corresponding region in space to have sufficient excitatory support to compete effectively with other regions.

The two models make distinct predictions (Cohen et al., 1994; O'Reilly & Munakata, 2000). For example, following *bilateral* parietal damage, the disengage model predicts disengage deficits on both sides of space, but the competitive inhibition model predicts *reduced* attentional effects (smaller valid and invalid trial effects). Data support the latter model (Coslett & Saffran, 1991; Verfaellie, Rapcsak, & Heilman, 1990), demonstrating the utility of biologically-based computational models for alternative theories of cognitive phenomena.

Mechanisms of Memory

In a computer, there are several different kinds of memory systems, each specialized to optimize some characteristics at the cost of others. For example, the RAM in a system is much faster than hard disk memory, but it also has a much smaller capacity. There are basic tradeoffs between speed and capacity that are resolved by having different systems optimized separately for each. Interestingly, human memory can also be understood in terms of a set of tradeoffs between different incompatible capacities. These basic tradeoffs are different than those behind the computer components (although one can see some similarities) — they are motivated instead by a consideration of conflicting capacities of neural networks. We discuss two different kinds of tradeoffs here, one that can help us understand the complementary roles of the hippocampus and cortex in learning, and another that relates to the specializations of the frontal cortex in working memory.

Complementary Hippocampal and Cortical Learning Systems

One important set of tradeoffs involves two basic types of learning that an organism must engage in — learning about specifics versus learning about generalities (figure 11). Because the neural mechanisms for achieving these types of learning are in direct conflict, the brain has evolved two separate brain structures to achieve these types of learning (McClelland, McNaughton, & O'Reilly, 1995; O'Reilly & Rudy, in press, 2000). The hippocampus appears to be specialized for learning about specifics, while the neocortex is good at extracting generalities.

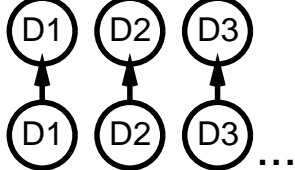
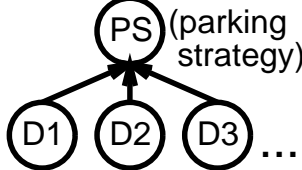
Two Incompatible Goals		
	Remember Specifics	Extract Generalities
Example:	Where is car parked?	Best parking strategy?
Need to:	Avoid interference	Accumulate experience
<i>Solution:</i>		
1.	Separate reps (keep days separate) 	Overlapping reps (integrate over days) 
2.	Fast learning (encode immediately)	Slow learning (integrate over days)
3.	Learn automatically (encode everything)	Task-driven learning (extract relevant stuff)
<i>These are incompatible, need two different systems:</i>		
System:	Hippocampus	Neocortex

Figure 11: Computational motivation for two complementary learning & memory systems in the brain, because there are two incompatible goals that such systems need to solve. One goal is to remember specific information, in this example where one's car is parked on a specific day. The other goal is to extract generalities across many experiences, for example in developing the best parking strategy over a number of different experiences. The neural solutions to these goals are incompatible: one requires representations to be kept separate, learned quickly, and automatically, while the other requires overlapping representations and slow learning to integrate over experiences, and is driven by task-specific constraints. Thus, it makes sense to have two separate neural systems separately optimized for each of these goals.

Specifically, learning about specifics requires keeping representations separated (to avoid interference), whereas learning about generalities requires overlapping representations that encode shared structure across many different experiences. Furthermore, learning about generalities requires a slow learning rate to gradually integrate new information with existing knowledge, while learning about specifics can occur rapidly. This rapid learning is particularly important for *episodic* memory, where the goal is to encode the details of specific events as they unfold.

In the example in figure 11, one can encode different kinds of information from experiences related to parking one's car. If one wants to remember the specifics of where the car is parked on a given day, it is important to encode this information using representations (populations of neurons) that are separate from representations for other such events, to minimize the *interference* that leads to forgetting. In a neural network, interference results from weights shared across multiple representations, because the different representations will pull these weights in different directions. Furthermore, one has only a short period of time to encode the parking location (unless you want to sit there and study it for hours), so rapid learning is required.

In contrast, if one wants to learn about the best strategy for parking (e.g., best location for a given time of day), one needs to integrate over many different experiences because any given day's experience does not provide a statistically reliable picture of the average situation. To accumulate information over individual experiences, one needs to ensure that these different experiences affect at least some of the same underlying neural representations — if you want to add things up, you need to put them all in the same place. Furthermore, given that the goal is computing something like an average, each event needs to make a relatively small contribution. In computing an average, you multiply each number by $\frac{1}{N}$, where N is the total number of items (events) to average over — as this becomes larger, each event makes a smaller contribution. In neural terms, this means using a small learning rate so that weights change only a small amount for each experience.

Thus, it is clear that these two kinds of learning are in direct conflict, and therefore that it would make sense to have two different neural systems specialized for each of these types of learning. This conclusion coincides nicely with a large body of data concerning the properties of the hippocampus and the cortex. It has been known for some time that damage to the hippocampus in the medial temporal lobe can produce severe memory deficits, while also leaving unimpaired certain kinds of learning and memory (Scoville & Milner, 1957; Squire,

1992). Although the precise characterization of the contributions of the hippocampus versus surrounding cortical areas has been a topic of considerable debate, it is possible to reconcile much of the data with the computational principles just described (O'Reilly & Rudy, in press). Furthermore, detailed biological properties of the hippocampus are ideally suited for maximizing the separation between neural representations of different events, enabling rapid episodic learning with minimal interference (O'Reilly & McClelland, 1994).

In the domain of human memory, the dual mechanisms of neocortex and hippocampus provide a natural fit with dual-process models of recognition memory (Jacoby, Yonelinas, & Jennings, 1997; Aggleton & Shaw, 1996; Aggleton & Brown, 1999; Vargha-Khadem, Gadian, Watkins, Connelly, Van Paesschen, & Mishkin, 1997; Holdstock, Mayes, Roberts, Cezayirli, Isaac, O'Reilly, & Norman, in press; Curran, 2000; O'Reilly, Norman, & McClelland, 1998). These models hold that recognition can be subserved by two different processes, a *recollection* process and a *familiarity* process. Recollection involves the recall of specific episodic details about the item, and thus fits well with the hippocampal principles developed here. Indeed, we have simulated distinctive aspects of recollection using a model based on many of the detailed biological properties of the hippocampus (O'Reilly et al., 1998). Familiarity is a non-specific sense that the item has been seen recently — we argue that this can be subserved by the small weight changes produced by slow cortical learning. Current simulation work has shown that a simple cortical model can account for a number of distinctive properties of the familiarity signal (Norman, O'Reilly, & Huber, 2000).

Models implementing the specialized hippocampal and cortical systems have also been shown to account for a wide range of learning and memory findings in rats, including nonlinear discrimination, incidental conjunctive encoding, fear conditioning, and transitive inference (O'Reilly & Rudy, in press). Also, there are a large number of important models of the hippocampus and/or cortical learning systems in the literature, many of which share important features with those described here (e.g., Marr, 1971; Treves & Rolls, 1994; Hasselmo & Wyble, 1997; Moll & Miikkulainen, 1997; Alvarez & Squire, 1994; Levy, 1989; Burgess, Recce, & O'Keefe, 1994; Samsonovich & McNaughton, 1997).

Complementary Posterior and Prefrontal Cortical Systems

Another important set of tradeoffs involves the extent to which a representation activates related representations, for example, the extent to which a neural representation of "smoke" activates the associated representa-

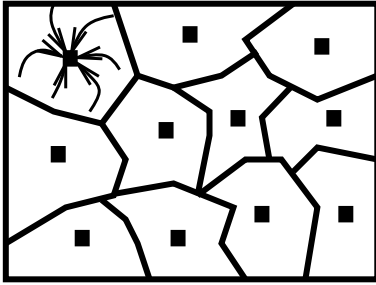


Figure 12: Attractor states (small squares) and their basins of attraction (surrounding regions), where nearby activation states are attracted to the central attractor state. Each stable attractor state could be used to actively maintain information over time. Note that the two-dimensional activation space represented here is a considerable simplification of the high-dimensional activation state over all the units in the network.

tion of “fire.” In some cases, such as when you want to remember that it was actually only smoke that you saw coming from the forest and not fire (e.g., to provide an accurate report about the situation to others), it would be best to actively maintain only smoke without activating fire. In other cases, such as when you want to form inferences based on seeing the smoke (e.g., to evaluate possible courses of action to take, such as bringing water), it would be best for smoke to activate fire. These goals of activating versus not activating related representations are obviously in conflict (and this problem gets much worse when the inferences are less certain than smoke \rightarrow fire); this tradeoff provides a potential way to understand the specializations between posterior cortex and prefrontal cortex. Specifically, prefrontal cortex may be specialized for active maintenance (a component of working memory) without activating associated representations, whereas posterior cortex may be specialized for inference based on activating associated representations.

The most obvious neural network mechanism for achieving active maintenance is recurrent bidirectional excitatory connectivity, where activation constantly circulates among active units, refreshing and maintaining their activation (Braver, Cohen, & Servan-Schreiber, 1995; Dehaene & Changeux, 1989; Munakata, 1998; Zipser, Kehoe, Littlewort, & Fuster, 1993). One can think of the effects of these recurrent connections in terms of an *attractor*, where the activation pattern of the network is attracted toward a stable state that persists over time (figure 12). An attractor is useful for memory because any perturbation away from that activation state is pulled back into the attractor, allowing in principle for relatively robust active maintenance in the face of noise and interference from ongoing processing.

The area around the attractor where perturbations are

pulled back is called the *basin of attraction*. For robust active maintenance, one needs to have attractors with wide basins of attraction, so that noise and other sources of interference will not pull the network out of its attractor. When there are many closely related representations linked by distributed connections, the basin of attraction around each representation is relatively narrow (i.e., the network can easily slip from one representation into the next). Thus, densely interconnected distributed representations will tend to conflict with the ability to maintain a specific representation actively over time.

The prefrontal cortex appears to have the relevant specializations for active maintenance. There is considerable physiological evidence that the prefrontal cortex subserves the active maintenance of information over time (i.e., as encoded in the persistent firing of frontal neurons) (e.g., Fuster, 1989; Goldman-Rakic, 1987; Miller, Erickson, & Desimone, 1996). Many computational models of this basic active maintenance function have been developed (Braver et al., 1995; Dehaene & Changeux, 1989; Zipser et al., 1993; Seung, 1998; Durstewitz, Seamans, & Sejnowski, 2000; Camperi & Wang, 1997). Further, the prefrontal cortex may have more isolated patterns of connectivity — neurons there appear to be interconnected within self-contained “stripe” patterns (Levitt, Lewis, Yoshioka, & Lund, 1993), and iso-coding microcolumns of neurons have been recorded (Rao, Williams, & Goldman-Rakic, 1999). Computational models have explored the impact of such connectivity and attractor dynamics on active maintenance (O’Reilly, Braver, & Cohen, 1999a; O’Reilly, Mozer, Munakata, & Miyake, 1999b). Models in which there are features that can participate equally in different distributed representations effectively have no attractors, and cannot maintain information over time in the absence of external inputs. The activation instead spreads across the distributed representations, resulting in a loss of the original information. With distributed representations that sustain attractors, active maintenance succeeds, but not in the presence of significant amounts of noise — wider attractor basins are necessary. With such wider attractor basins, as when units are completely *isolated* from each other, this completely prevents activation spread and yields very robust active maintenance, but at the loss of the ability to perform inference via activation spread.

Thus, computational models and considerations have helped to understand the specializations of posterior and prefrontal cortex, and how the prefrontal cortex might play a role in subserving working memory.

The Prefrontal Cortex and Higher Level Cognition

The prefrontal cortex is also important for a range of complex cognitive functions, such as planning and problem solving, described generally as falling under the umbrella of *higher level cognition*. Many theories summarize the function of frontal cortex in terms of “executive control,” “controlled processing,” or a “central executive” (e.g., Baddeley, 1986; Shallice, 1982; Gathercole, 1994; Shiffrin & Schneider, 1977), without explaining at a mechanistic level how such functionality could be achieved or why the prefrontal cortex would be specialized for such functionality. We saw in the preceding section how a consideration of computational tradeoffs has helped to understand the issue of specialization. In this section, we see how computational models have provided an important tool for exploring specific mechanisms that might achieve executive-like functionality.

One proposal along these lines is that the fundamental mechanism of active maintenance enables all the other executive-like functionality ascribed to the frontal cortex (Cohen, Braver, & O’Reilly, 1996; Goldman-Rakic, 1987; Munakata, 1998; O’Reilly et al., 1999a; O’Reilly & Munakata, 2000; Roberts & Pennington, 1996). As elaborated below, a number of models have demonstrated that active maintenance can account for frontal involvement in a range of different tasks that might otherwise appear to have nothing to do with simply maintaining information over time.

For example, several models have demonstrated that frontal contributions to “inhibitory” tasks can be explained in terms of active maintenance instead of an explicit inhibitory function. Actively-maintained representations can support (via bidirectional excitatory connectivity) correct choices, which will therefore indirectly inhibit incorrect ones via standard lateral inhibition mechanisms within the cortex. A model of the Stroop task provided an early demonstration of this point (Cohen, Dunbar, & McClelland, 1990). In this task, color words (e.g., “red”) are presented in different colors, and people are instructed to either read the word or name the color of ink that the word is written in. In the conflict condition, the ink color and word are different. Because we have so much experience reading, we naturally tend to read the word, even if instructed to name the color, such that responses are slower and more error-prone in the color-naming conflict condition than the word-reading one. These color-naming problems are selectively magnified with frontal damage. This frontal deficit has typically been interpreted in terms of the frontal cortex helping to inhibit the dominant word-reading pathway. However, Cohen et al. (1990) showed that they

could account for both normal and frontal-damage data by assuming that the frontal cortex instead supports the color-naming pathway, which then collaterally inhibits the word-reading pathway. Similar models have demonstrated that in infants, the ability to inhibit *perseverative* reaching (searching for a hidden toy at a previous hiding location rather than at its current location) can develop simply through increasing abilities to actively maintain a representation of the correct hiding location (Dehaene & Changeux, 1989; Munakata, 1998). Again, such findings challenge the standard interpretation that inhibitory abilities *per se* must develop for improved performance on this task (Diamond, 1991).

The activation-based processing model of frontal function can also explain why frontal cortex facilitates rapid switching between different categorization rules in the Wisconsin card sorting task and related tasks. In these tasks, subjects learn to categorize stimuli according to one rule via feedback from the experimenter, and then the rule is switched. With frontal damage, patients tend to perseverate in using the previous rule. A computational model of a related ID/ED categorization task demonstrated that the ability to rapidly update active memories in frontal cortex can account for detailed patterns of data in monkeys with frontal damage (O’Reilly, Noelle, Braver, & Cohen, submitted; O’Reilly & Munakata, 2000).

In short, computational models of frontal function can provide mechanistic explanations that unify the disparate roles of the frontal cortex, from working memory to cognitive control and planning/problem solving. However, a major remaining challenge is to explore whether truly complex “intelligent” behavior can be captured using these basic mechanisms.

Challenges

Most researchers agree that *if* a network model captures in sufficient detail the essential neural processes, then it can provide a truly valuable tool for advancing our understanding of the relation between brain and mind. However, there is skepticism regarding whether (a) enough is known about the neurobiology at this time to sufficiently constrain models, and (b) current models violate or fail to include important aspects of the known neurobiology.

We contrasted errors of omission (aspects of the biology that are missing or simplified in the models) with errors of commission (aspects of the models that are unlikely to be true given what we already know about the brain). We saw that in many cases, network models make errors of omission, but not errors of commission. For example, it is possible to make network models that

make no errors of commission at the level of network interactions, in that they follow the general excitatory and inhibitory connectivity patterns of the cortex (e.g., O'Reilly & Munakata, 2000; Somers, Nelson, & Sur, 1995; Lumer, Edelman, & Tononi, 1997 and many others). However, these networks undoubtedly make many errors of omission, given that there is considerable complexity in the wiring structures of the human cortex. As in other cases discussed above, it is not yet clear what functional significance (if any) these omissions might have.

In the few cases where there may be errors of commission (e.g., in error-driven learning algorithms), strong top-down constraints from cognition (e.g., the fact that people can learn difficult tasks) drive these possibly problematic properties. Considerable progress has been made in developing error-driven learning algorithms that are more consistent with known biology (while retaining the powerful capabilities), but there are several assumptions that remain untested.

Instead of denying outright the value of any given approach, we argue that science will be advanced through the contest of different theories as they attempt to explain increasing amounts of data, and that computational models provide a valuable source of theorizing that can provide novel insights and approaches to understanding complex, cognitive neuroscience phenomena. This is true even if the models are simplified and even if they contain some aspects that violate what we know about the brain — verbal theories are equally (if not more) likely to contain the same flaws. Often, however, these flaws are hidden by the vagueness of the verbal theories, while computational models have the virtue/vice of exposing all the gory details and assumptions required to actually implement a working simulation.

In short, we think that a major value of computational modeling is engaging in the *process* of working out explicit, mechanistic theories of how the brain gives rise to cognitive function. This process is iterative, cumulative, and not without controversy. However, its primary advantage is in directly confronting the major questions that need to be answered to understand how the brain does what it does.

General Discussion

In this article, we have touched on most of the central aspects of computational neural network models for psychological modeling. Building from individual neurons to networks thereof, we have shown how these networks incorporate many detailed aspects of the known neurobiology, while still remaining somewhat abstract. We emphasized that there are modeling formalisms that do

not make any obvious errors of commission — they do not violate any well known properties of the neural networks of the brain. Nevertheless, it remains to be tested how important the many errors of omission are for the biological fidelity of these models. We then showed how these models can speak to important issues in cognitive neuroscience, including issues in perception, attention, memory, and higher level cognition.

In the domain of perception, we showed how basic learning mechanisms and forms of neural interaction (inhibitory competition) can lead to the development of efficient representations for encoding the visual environment. We further summarized how attentional effects, which are needed to manage the overflow of perceptual input, fall naturally out of the combined neural dynamics of bidirectional connectivity and inhibitory competition. When these neural mechanisms are used to simulate spatial attention tasks widely used in cognitive psychology, they provide novel explanations of both intact and brain damaged performance, which accord better with the data than other theories based on a more abstract information-processing approach.

In the domain of learning and memory, we showed how an understanding of the capacities of fundamental neural mechanisms can lead to insights into how the brain has divided up the overall function of memory. Specifically, computational tradeoffs — between learning specifics versus learning generalities and between interconnected and isolated representations — suggest that different brain areas should be specialized to perform these different functions. This fits well with a wide range of data. Thus, the computational models help us to understand not only *how* the brain is organized to perform cognitive functions, but also *why* it might be organized this way in the first place.

In the domain of higher level cognition, we showed how models have helped to begin addressing the mechanisms that might underlie complex behaviors, such as those that require moving beyond habitual or prepotent responses. Specifically, active maintenance subserved by prefrontal cortex may support alternative choices, allowing habitual behaviors to be inhibited via lateral inhibitory mechanisms within the cortex. The ability to rapidly update activation-based representations in prefrontal cortex may be a critical component of flexible behavior.

In conclusion, we hope these examples provide a sufficient basis to understand both the strengths of neural network models and the criticisms surrounding them. Even though there are undoubtedly many missing features of these models, we think they capture enough of the most important properties to provide satisfying simulations of cognitive phenomena. Furthermore, the very

endeavor of creating these models raises a large number of important questions that are only beginning to be answered. Models should thus serve as an important part of the process of scientific progress in understanding human cognition.

References

- Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science*, 9, 147–169.
- Aggleton, J. P., & Brown, M. W. (1999). Episodic memory, amnesia, and the hippocampal-anterior thalamic axis. *Behavioral and Brain Sciences*, 22, 425–490.
- Aggleton, J. P., & Shaw, C. (1996). Amnesia and recognition memory: A re-analysis of psychometric data. *Neuropsychologia*, 34, 51–62.
- Alvarez, P., & Squire, L. R. (1994). Memory consolidation and the medial temporal lobe: A simple network model. *Proceedings of the National Academy of Sciences, USA*, 91, 7041–7045.
- Baddeley, A. D. (1986). *Working memory*. New York: Oxford University Press.
- Bear, M. F., & Malenka, R. C. (1994). Synaptic plasticity: LTP and LTD. *Current Opinion in Neurobiology*, 4, 389–399.
- Bell, A. J., & Sejnowski, T. J. (1997). The independent components of natural images are edge filters. *Vision Research*, 37, 3327–3338.
- Bienenstock, E. L., Cooper, L. N., & Munro, P. W. (1982). Theory for the development of neuron selectivity: Orientation specificity and binocular interaction in visual cortex. *Journal of Neuroscience*, 2(2), 32–48.
- Braver, T. S., Cohen, J. D., & Servan-Schreiber, D. (1995). A computational model of prefrontal cortex function. In D. S. Touretzky, G. Tesauro, & T. K. Leen (Eds.), *Advances in neural information processing systems* (pp. 141–148). Cambridge, MA: MIT Press.
- Bryson, A. E., & Ho, Y. C. (1969). *Applied optimal control*. New York: Blaisdell.
- Burgess, N., Recce, M., & O'Keefe, J. (1994). A model of hippocampal function. *Neural networks*, 7, 1065–1083.
- Camperi, M., & Wang, X. J. (1997). Modeling delay-period activity in the prefrontal cortex during working memory tasks. In J. Bower (Ed.), *Computational neuroscience* (Chap. 44, pp. 273–279). New York: Plenum Press.
- Carpenter, G., & Grossberg, S. (1987). A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics, and Image Processing*, 37, 54–115.
- Cohen, J. D., Braver, T. S., & O'Reilly, R. C. (1996). A computational approach to prefrontal cortex, cognitive control, and schizophrenia: Recent developments and current challenges. *Philosophical Transactions of the Royal Society (London) B*, 351, 1515–1527.
- Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing model of the Stroop effect. *Psychological Review*, 97(3), 332–361.
- Cohen, J. D., Romero, R. D., Farah, M. J., & Servan-Schreiber, D. (1994). Mechanisms of spatial attention: The relation of macrostructure to microstructure in parietal neglect. *Journal of Cognitive Neuroscience*, 6, 377.
- Coslett, H. B., & Saffran, E. (1991). Simultanagnosia. To see but not two see. *Brain*, 114, 1523–1545.
- Crick, F. H. C. (1989). The recent excitement about neural networks. *Nature*, 337, 129–132.
- Curran, T. (2000). Brain potentials of recollection and familiarity. *Memory and Cognition*, 28, 923.
- Dehaene, S., & Changeux, J. P. (1989). A simple model of prefrontal cortex function in delayed-response tasks. *Journal of Cognitive Neuroscience*, 1, 244–261.
- Diamond, A. (1991). Neuropsychological insights into the meaning of object concept development. In S. Carey, & R. Gelman (Eds.), *The epigenesis of mind* (Chap. 3, pp. 67–110). Mahwah, NJ: Lawrence Erlbaum.
- Durstewitz, D., Seamans, J. K., & Sejnowski, T. J. (2000). Dopamine-mediated stabilization of delay-period activity in a network model of prefrontal cortex. *Journal of Neurophysiology*, 83, 1733.
- Edelman, G. (1987). *Neural Darwinism*. New York: Basic Books.
- 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.
- Erwin, E., Obermayer, K., & Schulten, K. (1995). Models of orientation and ocular dominance columns in the visual cortex: A critical comparison. *Neural Computation*, 7, 425–468.
- Fuster, J. M. (1989). *The prefrontal cortex: Anatomy, physiology and neuropsychology of the frontal lobe*. New York: Raven Press.
- Gathercole, S. E. (1994). Neuropsychology and working memory: A review. *Neuropsychology*, 8(4), 494–505.
- Goldman-Rakic, P. S. (1987). Circuitry of primate prefrontal cortex and regulation of behavior by representational memory. *Handbook of Physiology — The Nervous System*, 5, 373–417.

- Grossberg, S. (1976). Adaptive pattern classification and universal recoding I: Parallel development and coding of neural feature detectors. *Biological Cybernetics*, 23, 121–134.
- Hasselmo, M. E., & Wyble, B. (1997). Free recall and recognition in a network model of the hippocampus: Simulating effects of scopolamine on human memory function. *Behavioural Brain Research*, 89, 1–34.
- Hebb, D. O. (1949). *The organization of behavior*. New York: Wiley.
- Hinton, G. E. (1989). Deterministic Boltzmann learning performs steepest descent in weight-space. *Neural Computation*, 1, 143–150.
- Hinton, G. E., & McClelland, J. L. (1988). Learning representations by recirculation. In D. Z. Anderson (Ed.), *Neural Information Processing Systems, 1987* (pp. 358–366). New York: American Institute of Physics.
- Holdstock, J. S., Mayes, A. R., Roberts, N., Cezayirli, E., Isaac, C. L., O'Reilly, R. C., & Norman, K. A. (in press). Memory dissociations following human hippocampal damage. *Hippocampus*.
- Hubel, D., & Wiesel, T. N. (1962). Receptive fields, binocular interaction, and functional architecture in the cat's visual cortex. *Journal of Physiology*, 160, 106–154.
- Jacoby, L. L., Yonelinas, A. P., & Jennings, J. M. (1997). The relation between conscious and unconscious (automatic) influences: A declaration of independence. In J. D. Cohen, & J. W. Schooler (Eds.), *Scientific approaches to consciousness* (pp. 13–47). Mahway, NJ: Lawrence Erlbaum Associates.
- Jaffe, D. B., & Carnevale, N. T. (1999). Passive normalization of synaptic integration influenced by dendritic architecture. *Journal of Neurophysiology*, 82, 3268–3285.
- Koch, C., & Segev, I. (Eds.). (1998). *Methods in neuronal modeling, 2d ed.* Cambridge, MA: MIT Press.
- Kohonen, T. (1984). *Self-organization and associative memory*. Berlin: Springer Verlag.
- Levitt, J. B., Lewis, D. A., Yoshioka, T., & Lund, J. S. (1993). Topography of pyramidal neuron intrinsic connections in macaque monkey prefrontal cortex (areas 9 & 46). *Journal of Comparative Neurology*, 338, 360–376.
- Levy, W. B. (1989). A computational approach to hippocampal function. In R. D. Hawkins, & G. H. Bower (Eds.), *Computational models of learning in simple neural systems* (pp. 243–304). San Diego, CA: Academic Press.
- Linsker, R. (1988). Self-organization in a perceptual network. *Computer*, 21(3), 105–117.
- Lisman, J. (1994). The CaM Kinase II hypothesis for the storage of synaptic memory. *Trends in Neurosciences*, 17, 406.
- Lisman, J. E. (1989). A mechanism for the Hebb and the anti-Hebb processes underlying learning and memory. *Proceedings of the National Academy of Sciences*, 86, 9574–9578.
- Livingstone, M., & Hubel, D. (1988). Segregation of form, color, movement, and depth: Anatomy, physiology, and perception. *Science*, 240, 740–749.
- Lumer, E., Edelman, G., & Tononi, G. (1997). Neural dynamics in a model of the thalamocortical system I. layers, loops and the emergence of fast synchronous rhythms. *Cerebral Cortex*, 7, 207–227.
- Marr, D. (1971). Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society (London) B*, 262, 23–81.
- Marr, D. (1982). *Vision*. New York: Freeman.
- McClelland, J. L. (1994). The interaction of nature and nurture in development: A parallel distributed processing perspective. In P. Bertelson, P. Eelen, & G. D'Ydewalle (Eds.), *Current advances in psychological science: Ongoing research* (pp. 57–88). Hillsdale, NJ: Erlbaum.
- 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., & Rumelhart, D. E. (Eds.). (1988). *Explorations in parallel distributed processing: A handbook of models, programs, and exercises*. Cambridge, MA: MIT Press.
- Miller, E. K., Erickson, C. A., & Desimone, R. (1996). Neural mechanisms of visual working memory in prefrontal cortex of the macaque. *Journal of Neuroscience*, 16, 5154.
- Miller, K. D., Keller, J. B., & Stryker, M. P. (1989). Ocular dominance column development: Analysis and simulation. *Science*, 245, 605–615.
- Minsky, M. L., & Papert, S. A. (1969). *Perceptrons*. Cambridge, MA: MIT Press.
- Moll, M., & Miikkulainen, R. (1997). Convergence-zone episodic memory: Analysis and simulations. *Neural Networks*, 10, 1017–1036.
- Movellan, J. R. (1990). Contrastive Hebbian learning in the continuous Hopfield model. In D. S. Touretzky,

- G. E. Hinton, & T. J. Sejnowski (Eds.), *Proceedings of the 1989 Connectionist Models Summer School* (pp. 10–17). San Mateo, CA: Morgan Kaufman.
- Munakata, Y. (1998). Infant perseveration and implications for object permanence theories: A PDP model of the \overline{AB} task. *Developmental Science*, *1*, 161–184.
- Munakata, Y., & Stedron, J. M. (in press). Neural network models of cognitive development. In C. A. Nelson, & M. Luciana (Eds.), *Handbook of developmental cognitive neuroscience*. Cambridge, MA: MIT Press.
- Norman, K. A., O'Reilly, R. C., & Huber, D. E. (2000). Modeling neocortical contributions to recognition memory. *The Cognitive Neuroscience Meeting, 2000*.
- Oja, E. (1982). A simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, *15*, 267–273.
- Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*, 607.
- O'Reilly, R. C. (1996a). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation*, *8*(5), 895–938.
- O'Reilly, R. C. (1996b). *The Leabra model of neural interactions and learning in the neocortex*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA.
- O'Reilly, R. C. (1998). Six principles for biologically-based computational models of cortical cognition. *Trends in Cognitive Sciences*, *2*(11), 455–462.
- O'Reilly, R. C., Braver, T. S., & Cohen, J. D. (1999a). 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*. (pp. 375–411). New York: Cambridge University Press.
- O'Reilly, R. C., & McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall: Avoiding a tradeoff. *Hippocampus*, *4*(6), 661–682.
- O'Reilly, R. C., Mozer, M., Munakata, Y., & Miyake, A. (1999b). Discrete representations in working memory: A hypothesis and computational investigations. *The Second International Conference on Cognitive Science* (pp. 183–188). Tokyo: Japanese Cognitive Science Society.
- O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. Cambridge, MA: MIT Press.
- O'Reilly, R. C., Noelle, D., Braver, T. S., & Cohen, J. D. (submitted). Prefrontal cortex and dynamic categorization tasks: Representational organization and neuromodulatory control.
- O'Reilly, R. C., Norman, K. A., & McClelland, J. L. (1998). A hippocampal model of recognition memory. In M. I. Jordan, M. J. Kearns, & S. A. Solla (Eds.), *Advances in neural information processing systems 10* (pp. 73–79). Cambridge, MA: MIT Press.
- O'Reilly, R. C., & Rudy, J. W. (2000). Computational principles of learning in the neocortex and hippocampus. *Hippocampus*, *10*, 389–397.
- O'Reilly, R. C., & Rudy, J. W. (in press). Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psychological Review*.
- Parker, D. B. (1985). *Learning logic* (Technical Report TR-47). Cambridge, MA: Center for Computational Research in Economics and Management Science, Massachusetts Institute of Technology.
- Peterson, C., & Anderson, J. R. (1987). A mean field theory learning algorithm for neural networks. *Complex Systems*, *1*, 995–1019.
- Posner, M. I., Walker, J. A., Friedrich, F. J., & Rafal, R. D. (1984). Effects of parietal lobe injury on covert orienting of visual attention. *Journal of Neuroscience*, *4*, 1863–1874.
- Rao, S. G., Williams, G. V., & Goldman-Rakic, P. S. (1999). Isodirectional tuning of adjacent interneurons and pyramidal cells during working memory: Evidence for microcolumnar organization in PFC. *Journal of Neurophysiology*, *81*, 1903.
- Reike, F., Warland, D., van Steveninck, R., & Bialek, W. (1996). *Spikes: Exploring the neural code*. Cambridge, MA: MIT Press.
- Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variation in the effectiveness of reinforcement and non-reinforcement. In A. H. Black, & W. F. Prokasy (Eds.), *Classical conditioning II: Theory and research* (pp. 64–99). New York: Appleton-Century-Crofts.
- Roberts, R. J., & Pennington, B. F. (1996). An interactive framework for examining prefrontal cognitive processes. *Developmental Neuropsychology*, *12*(1), 105–126.
- Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & PDP Research Group (Eds.), *Parallel distributed processing. Volume 1: Foundations* (Chap. 8, pp. 318–362). Cambridge, MA: MIT Press.

- Rumelhart, D. E., & Zipser, D. (1986). Feature discovery by competitive learning. In D. E. Rumelhart, J. L. McClelland, & PDP Research Group (Eds.), *Parallel distributed processing. Volume 1: Foundations* (Chap. 5, pp. 151–193). Cambridge, MA: MIT Press.
- Samsonovich, A., & McNaughton, B. L. (1997). Path integration and cognitive mapping in a continuous attractor neural network model. *Journal of Neuroscience*, *17*, 5900–5920.
- Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry*, *20*, 11–21.
- Seung, H. S. (1998). Continuous attractors and oculomotor control. *Neural Networks*, *11*, 1253.
- Shallice, T. (1982). Specific impairments of planning. *Philosophical Transactions of the Royal Society (London) B*, *298*, 199–209.
- Shepherd, G. M. (Ed.). (1990). *The synaptic organization of the brain*. Oxford: Oxford University Press.
- Shepherd, G. M., & Brayton, R. K. (1987). Logic operations are properties of computer-simulated interactions between excitable dendritic spines. *Neuroscience*, *21*, 151–166.
- Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending, and a general theory. *Psychological Review*, *84*, 127–190.
- Somers, D., Nelson, S., & Sur, M. (1995). An emergent model of orientation selectivity in cat visual cortical simple cells. *Journal of Neuroscience*, *15*, 5448.
- Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review*, *99*, 195–231.
- Swindale, N. V. (1996). The development of topography in the visual cortex: A review of models. *Network: Computation in Neural Systems*, *7*, 161–247.
- Tovee, M. J., Rolls, E. T., Treves, A., & Bellis, R. P. (1993). Information encoding and the responses of single neurons in the primate temporal visual cortex. *Journal of Neurophysiology*, *70*, 640–654.
- Treves, A., & Rolls, E. T. (1994). A computational analysis of the role of the hippocampus in memory. *Hippocampus*, *4*, 374–392.
- van Hateren, J. H., & van der Schaaff, A. (1997). Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society, London, B*, *265*, 359–366.
- Vargha-Khadem, F., Gadian, D. G., Watkins, K. E., Connelly, A., Van Paesschen, W., & Mishkin, M. (1997). Differential effects of early hippocampal pathology on episodic and semantic memory. *Science*, *277*, 376–380.
- Verfaellie, M., Rapcsak, S. Z., & Heilman, K. M. (1990). Impaired shifting of attention in Balint's syndrome. *Brain and Cognition*, *12*, 195–204.
- Werbos, P. (1974). *Beyond regression: New tools for prediction and analysis in the behavioral sciences*. PhD thesis, Harvard University.
- Widrow, B., & Hoff, M. E. (1960). Adaptive switching circuits. *Institute of Radio Engineers, Western Electronic Show and Convention, Convention Record, Part 4* (pp. 96–104).
- Zipser, D., & Andersen, R. A. (1988). A backpropagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature*, *331*, 679–684.
- Zipser, D., Kehoe, B., Littlewort, G., & Fuster, J. (1993). A spiking network model of short-term active memory. *Journal of Neuroscience*, *13*, 3406–3420.