Encyclopedia of Cognitive Sciences

Computational Neuroscience: From Biology to Cognition
Article Ref. Code 51

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

29th July 2001

Headword: Computational Neuroscience and Cognitive Modeling

Keywords: neural networks, cognitive modeling, cognitive neuroscience, vision, memory

Contents list:

Article definition: Computational neuroscience involves the construction of explicit computational models that implement neural mechanisms to simulate cognitive functions such as perception, learning and memory, motor function, and language.

Introduction

This article describes computer models that simulate the neural networks of the brain, with the goal of understanding how cognitive functions (perception, memory, thinking, language, etc) arise from their neural basis. Many neural network models have been developed over the years, focused at many different levels of analysis from engineering to relatively low-level biology to cognition. Here, we consider models that try to span the gap between biology and cognition, such that they deal with real cognitive data, using mechanisms that are related to the underlying biology.

The relationship between cognitive and neural theories

Computational models provide an important tool for linking data across multiple levels of analysis. The cognitive implications of cellular and network properties of neurons are often not immediately apparent — there are simply too many factors at many different levels interacting in complex ways. As a result, trying to develop behavioral predictions that capture the complexity of the neural level can be like trying to predict the weather from a number of satellite measurements. A computational model, of the weather or of the brain, can help by formalizing information and relating it through complex, emergent dynamics. Cognitive properties can thus be understood as the product of a number of lower-level interactions, and neural properties can be understood in terms of their functional role in cognitive processes. Further, the effects of manipulations to lower-level interactions (e.g., through genetic knockouts or lesions) can be simulated and reconciled with the observed behavioral effects. Importantly, these simulations can make sense of much more subtle behavioral effects than the generic impairment of behavior on a cognitive task.

Although models thus have the potential to inform brain-behavior relations, they do not always do so. Models can be underconstrained by neural and behavioral data, and thus of questionable value in understanding how the brain actually subserves behavior. Moreover, models can be put forth as mere demonstrations that a behavior can be simulated, but this is insufficient for understanding why the models behave as they do. Thus, models must be evaluated in a balanced way for whether they advance understanding of specific phenomena, provide general principles, and make useful links between brain and behavior.

In this chapter, we review a number of neuroscience-based computational models of various cognitive phenomena, with an emphasis on the general principles embodied by these models and their implications for understanding the general nature of cognition. Specifically, we examine models of: vision, including topography and receptive fields in primary visual cortex and spatial attention emerging from interactions between parietal and temporal streams of processing; episodic memory subserved by the hippocampus; conditioning and skill learning subserved by the basal ganglia and cerebellum; working memory and cognitive control subserved by the prefrontal cortex; and language processing guided by neuropsychological cases. For a more comprehensive treatment of many of these models and the ideas behind them, see O'Reilly and Munakata (2000).

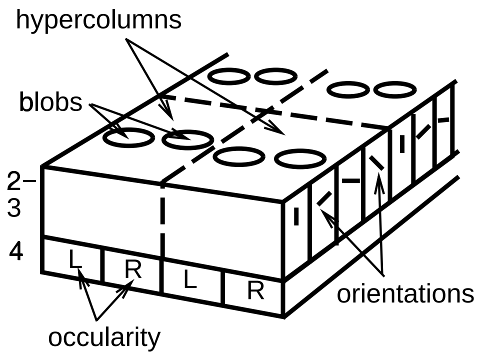

Figure 1: Structure of a cortical hypercolumn, that represents a full range of orientations (in layers 2–3), ocular dominance columns (in layer 4, one for each eye), and surface features (in the blobs). Each such hypercolumn is focused within one region of retinal space, and neighboring hypercolumns represent neighboring regions.

Computational models of vision guided by neuroscience

Vision is one of the best studied domains in cognitive neuroscience, having a long tradition of integrating biological and psychophysical levels of analysis. Computational models of vision have been influential in both the vision and computational communities. We review two areas of visual modeling here: topography and receptive fields in primary visual cortex (V1) and spatial attention and the effects of parietal lobe damage. Other major areas of visual processing that have been modeled include object recognition, motion processing, and figure-ground segmentation.

Topography and receptive fields in primary visual cortex

The primary visual cortex, V1, provides an interesting target for computational models, because it has a complex but relatively well-understood organization of visual feature detectors (a *representational structure*) subject to considerable experience-based developmental plasticity (Hubel & Wiesel, 1962; Gilbert, 1996). Thus, the overarching question behind many of the V1 models has been: *Can we reproduce the complex representational structure of V1 through principled learning mechanisms exposed to realistic visual inputs?*

First, we summarize the complex representational structure of V1. V1 neurons are generally described as *edge detectors*, where an edge is simply a roughly linear separation between a region of relative light and dark. These detectors differ in their orientation, size, position, and *polarity* (i.e., going from light-to-dark or dark-to-light, or dark-light-dark and light-dark-light). The different types of edge detectors (together with other neurons that appear to encode visual surface properties) are packed into the two-dimensional 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 1) containing smoothly varying progressions of oriented edge detectors, among other things (Livingstone & Hubel, 1988). The hypercolumn also contains *ocular dominance columns*, in which V1 neurons respond preferentially to input from one eye or the other.

Many computational models have emphasized one or a few aspects of the many detailed proper-

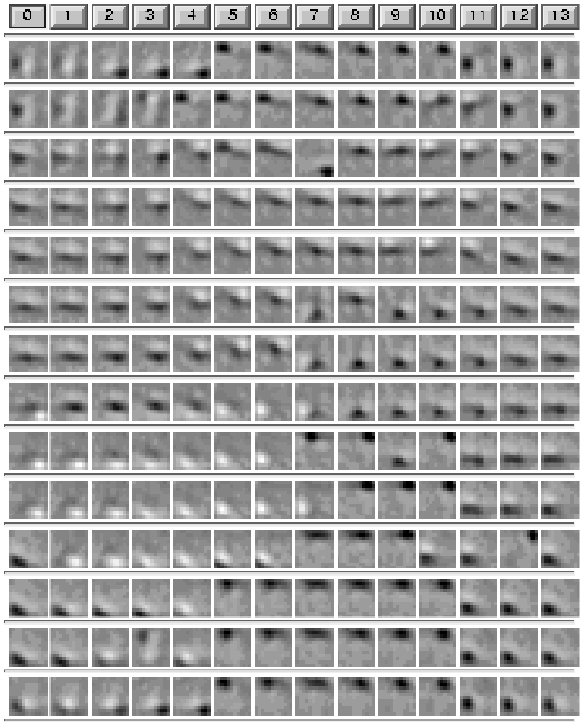

Figure 2: The receptive fields of model V1 neurons (from O'Reilly & Munakata, 2000). Lighter shades indicate areas of on-center response, and darker shades indicate areas of off-center response. Individual units are shown by smaller grids (showing weights into those units from different locations in the retinally-organized input) organized into a larger grid representing the location of each unit within the simulated V1 hypercolumn.

ties of V1 representations (for reviews, see Swindale, 1996; Erwin, Obermayer, & Schulten, 1995). For example, models have demonstrated how ocular dominance columns can develop based on a *Hebbian* learning mechanism, with greater local correlations in the neural firing coming from within one eye than from across eyes (Miller, Keller, & Stryker, 1989). Hebbian learning encodes correlational structure by strengthening the weights between neurons that fire together, and decreasing the weights between those that do not (see Oja, 1982; Linsker, 1988 for mathematical analyses of Hebbian correlational learning).

Several models have demonstrated how a realistic set of oriented edge-detector representations can develop in networks presented with natural visual scenes, 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 has shown how biologically-based models can develop oriented receptive fields, through a Hebbian learning mechanism with sparseness constraints in the form of inhibitory competition between neurons (a known property of cortex) (O'Reilly & Munakata, 2000). Furthermore, lateral excitatory connections within this network (another known property of cortex) produced a topographic organization consistent with several aspects of the hypercolumn structure (e.g., gradients of orientation, size, polarity, and phase tuning

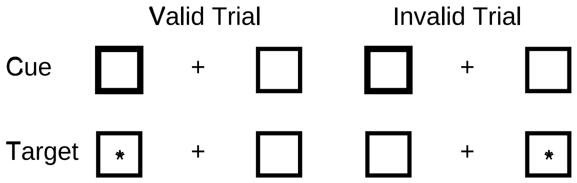

Figure 3: The Posner spatial attention task. The cue is a brightening or highlighting of one of the boxes that focuses attention to that region of space. Reaction times to detect the target are faster when this cue is valid (the target appears in that same region) than when it is invalid (the target appears elsewhere).

and pinwheel discontinuities; figure 2).

To summarize, these V1 models demonstrate how Hebbian learning mechanisms exposed to naturalistic stimuli, with certain kinds of biological prestructuring (e.g., connectivity patterns and inhibition), can produce aspects of the observed representational structure of V1. However, many complex aspects of early visual processing remain to be addressed, including motion, texture, and color sensitivity of different populations of V1 neurons.

Spatial attention and the effects of parietal lobe damage

Many computational models of higher-level vision have explored object recognition (e.g., Mozer, 1991; Fukushima, 1988; LeCun, Boser, Denker, Henderson, Howard, Hubbard, & Jackel, 1989) and spatial processing (e.g., Pouget & Sejnowski, 1997; Mozer & Sitton, 1998; Vecera & O'Reilly, 1998). Here we describe a model of spatial attention (Cohen, Romero, Farah, & Servan-Schreiber, 1994) that demonstrates how biologically-based computational models can provide alternative interpretations of cognitive phenomena. Spatial attention has classically been operationalized according to the Posner spatial cuing task (Posner, Walker, Friedrich, & Rafal, 1984, figure 3). 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.

According to the standard account of these data, spatial attention involves 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 recurrent excitatory connections and competitive inhibitory connections, provide an alternative explanation for these phenomena (Cohen et al., 1994; O'Reilly & Munakata, 2000). In this framework, the facilitory effects of drawing attention to one region of space result from excitatory connections between spatial and other representations of that region — this excitatory support makes it easier to process information in that region. The slowing that comes on the invalid trials results from inhibitory competition between different spatial regions. 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 (Posner et al., 1984), but the competitive inhibition model predicts *reduced* attentional effects (smaller valid and invalid trial effects). Data support the latter model (e.g., Coslett & Saffran, 1991; Verfaellie, Rapcsak, & Heilman, 1990), demonstrating the utility of biologically-based computational models for alternative theories of cognitive phenomena.

Computational models of episodic memory and the hippocampus

Damage to a brain structure called 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). The hippocampus has thus been a popular target of computational modeling to explore its exact contribution, and these models have had a large impact on the field (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).

One framework has combined known biological features of the hippocampal formation with computationally motivated principles about learning and memory to further clarify the unique contributions of the hippocampus in memory (McClelland, McNaughton, & O'Reilly, 1995; O'Reilly & Rudy, 2000, 2001; O'Reilly & McClelland, 1994; O'Reilly, Norman, & McClelland, 1998). The central idea is that there are two basic types of learning that an organism must engage in — learning about specifics versus learning about generalities — and that because the computational 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. The hippocampus appears to be specialized for learning about specifics, while the neocortex is good at extracting generalities.

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.

These computational principles provide a satisfying and precise characterization of the division of labor between the hippocampus and neocortex. The models that implement these principles have been shown to account for a wide range of specific learning and memory findings, including nonlinear discrimination, incidental conjunctive encoding, fear conditioning, and transitive inference in rats (O'Reilly & Rudy, 2001) and human recognition memory (O'Reilly et al., 1998). However, these models fail to incorporate important aspects of the hippocampal formation (e.g., the subiculum and the mossy cells in the hilus), and many more complex behaviors that depend on the hippocampus (and its interactions with other brain areas) remain to be addressed.

Computational models of conditioning and skill learning in the basal ganglia and cerebellum

A convergence between biological/behavioral and computational approaches has been achieved in the domain of conditioning (learning to associate stimuli/actions with rewards). In the computational domain, *reinforcement learning* mechanisms can adapt the behavior of a simulated animal according to reward contingencies in the environment (Sutton & Barto, 1998). Such learning mechanisms, including the *temporal differences* algorithm (Sutton, 1988), have been proven to not only work well mathematically (e.g., Dayan, 1992), but to also correspond with aspects of neural recordings made in the reward-processing area of the brain (Montague, Dayan, & Sejnowski, 1996; Schultz, Dayan, & Montague, 1997).

Specifically, a straightforward neural implementation of the temporal differences algorithm involves a systematic transition of reward-related neural firing similar to that observed in *dopamine* neurons in the midbrain. During a simple conditioning task where a sensory stimulus (e.g., a tone) reliably predicts a subsequent reward (e.g., juice), these neurons initially fire in response to the reward, but then after some trials of learning, they respond to the sensory stimulus that predicts the reward and no longer fire to the reward itself (Schultz, Apicella, & Ljungberg, 1993; Schultz, Apicella, Romo, & Scarnati, 1995). This transfer of reward-related firing from the actual reward to predictors of the reward is a key property of the temporal-differences mechanism as implemented by Montague, et al. (1996), which thus provides a principled, provably-effective explanation for why the brain appears to learn in this manner.

Models of motor performance and skill learning have been developed based on the biological properties of the relevant underlying brain areas including the basal ganglia (including the striatum, globus pallidus, substantia nigra, subthalamic nucleus, and the nucleus accumbens) and the cerebellum (e.g., Beiser, Hua, & Houk, 1997; Wickens, 1997; Houk, Davis, & Beiser, 1995; Berns & Sejnowski, 1996; Schweighofer, Arbib, & Kawato, 1998a, 1998b; Contreras-Vidal, Grossberg, & Bullock, 1997). These models make close contact with detailed neural properties of these areas, but tend to focus on simpler aspects of motor performance — complex motor skills remain to be addressed.

Computational models of working memory, cognitive control and prefrontal cortex

The prefrontal cortex is important for a range of cognitive functions that can be described generally as *higher level cognition*, in that they go beyond basic perceptual, motor, and memory functions. For example, frontal cortex has been implicated in problem solving tasks like the Tower of Hanoi/London (e.g., Shallice, 1982; Baker, Rogers, Owen, Frith, Dolan, Frackowiak, & Robbins, 1996; Goel & Grafman, 1995), which requires executing a sequence of moves to achieve a subsequent goal. Many theoretical perspectives 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. Computational models provide an important tool for exploring specific mechanisms that might achieve executive-like functionality.

Working memory and active maintenance

One proposal along these lines is that the fundamental mechanism underlying frontal function is *active maintenance,* which then enables all the other executive-like functionality ascribed to the frontal cortex (Cohen, Braver, & O'Reilly, 1996; Goldman-Rakic, 1987; Munakata, 1998; O'Reilly, Braver, & Cohen, 1999; O'Reilly & Munakata, 2000; Roberts & Pennington, 1996). For example, a flexible, adaptive active maintenance system can enable an entirely different kind of solution to information processing challenges — one that involves the strategic activation and de-activation of representations (*activation-based processing*) instead of weight changes (*weight-based processing*) (O'Reilly & Munakata, 2000). There are tradeoffs between these types of processing (e.g., activations can be more rapidly switched than weights, but they are also transient), so both kinds of processing are better than either alone.

There is considerable direct biological evidence that the frontal 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, Cohen, & Servan-Schreiber, 1995; Dehaene & Changeux, 1989; Zipser, Kehoe, Littlewort, & Fuster, 1993; Seung, 1998; Durstewitz, Seamans, & Sejnowski, 2000; Camperi & Wang, 1997). As elaborated below, a number of models have further 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.

Inhibition, flexibility, and perseveration

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

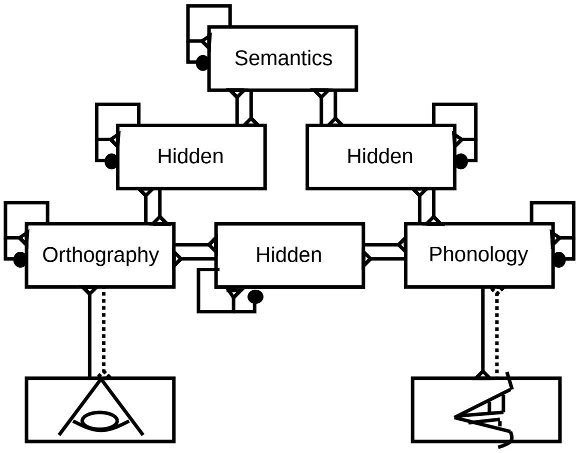

Figure 4: For the purposes of reading, words are represented in a distributed fashion across orthographic (visual word recognition), phonological (speech output), and semantic areas.

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, it remains to be shown whether truly complex "intelligent" behavior can be captured using these basic mechanisms.

Computational models of language use guided by neuropsychological cases

Damage to language-related brain areas causes a wide variety of impairments. One subset of such impairments, the *dyslexias* (also known as *alexias*) have been the subject of a series of influential computational models of the normal and impaired reading process (Seidenberg & McClelland, 1989; Plaut & Shallice, 1993; Plaut, McClelland, Seidenberg, & Patterson, 1996). These models simulate the pathways between visual word inputs (*orthography*), word semantics, and verbal word outputs (*phonology*), and can account for different kinds of dyslexias in terms of differential patterns of damage to these pathways (figure 4).

These models have been influential in part because they suggest an alternative, somewhat counterintuitive interpretation of how words are represented and how language processing works. Traditional models have assumed that the brain contains a *lexicon* with distinct representations for different words. Furthermore, these models assume that reading a word aloud (i.e., mapping between orthography and phonology) can occur via two different *routes*: pronunciation rules (for *regular* words like "make") or a lookup-table kind of mechanism (for *exception* words like "yacht") (Pinker, 1991; Coltheart, Curtis, Atkins, & Haller, 1993; Coltheart & Rastle, 1994). In contrast to these dual-route models, the neural network models allow for a single pathway to process both regular and exception words, and they employ a *distributed* lexicon without any centralized, discrete lexical representations. Instead, lexical processing occurs in pathways that map between different aspects of word representations (figure 4).

In general, neural networks can learn all kinds of different mappings — fully regular ones, like the spelling-sound mapping of the "a" in words like "make," "bake," and so on, as well as irregular mappings that occur in exceptions (e.g., "yacht"). Nevertheless, networks are sensitive to both the degree of regularity and the frequency of different mappings. Specifically, neural network models predict frequency-by-regularity interactions that would not be expected in dual-route models, and that are observed in behavioral tests (Plaut et al., 1996). Furthermore, these network models can account for patterns of deficit with brain damage that would seem improbable under dual-route models. For example, people with *surface dyslexia* can read nonwords (e.g., "nust"), but they are impaired at retrieving semantic information from written words, and have difficulty in reading exception words. Thus, it would be natural in neural network models to interpret this as damage in the pathway between orthography and semantics. Critically however, surface dyslexics' difficulty with exception words is generally limited to low-frequency exceptions (e.g., "yacht"; they can read high-frequency exceptions, e.g., "are"). This pattern suggests that the remaining "direct" pathway between orthography and phonology can handle both regulars and high-frequency exceptions, as is true of the network models. This pattern of data is not easily explained in the dual-route models — with two pathways, either regulars or exceptions should be affected, but not both, and not as a function of frequency.

In summary, neural network models of language can provide alternative, counterintuitive ways of explaining complex patterns of deficits that occur with brain damage. Nevertheless, this area remains highly controversial as neural network accounts are challenged by revised versions of dual-route models, and by the complexity of different neuropsychological profiles associated with damage to different language areas.

Summary

The above examples illustrate that computational models based on the neural networks of the brain can provide important insights, insights that might otherwise be difficult to obtain. Many models have applied a set of basic principles to a range of phenomena, and arrived at completely different explanations than those based on purely verbal cognitive theories. As a result, these models have played an important role in guiding empirical research and theorizing across a number of domains.

Despite these successes, many people remain skeptical of models. A common concern is that different models may employ different sets of mechanisms to explain the same data, such that it may not be that interesting when a given model can simulate a set of data. Several points have been made in response to this concern. First, this concern applies not only to computational models, but to scientific theorizing more generally (multiple competing theories can account for the same data), and the response is similar in each case (Munakata & Stedron, in press). Competing theories and models can be evaluated by many other criteria than simply accounting for a set of data, such as the accuracy of predictions, the coherence of the theoretical framework, and the ease of accounting for new data. Second, mechanisms developed independently can turn out to be equivalent (e.g., O'Reilly, 1996), providing converging evidence for their utility, and indicating more coherence to principles than might otherwise be evident. Third, a common set of mechanisms appears to be emerging as the field continues to mature. For example, over 40 different phenomena (including most of what was described above) have been modeled using a common set of mechanisms (O'Reilly & Munakata, 2000). This set of mechanisms was developed over many years by many different researchers, and has now been consolidated and integrated into one coherent framework (O'Reilly, 1998). Therefore, there is increasingly a largely consistent set of ideas underlying many neural network models, and this framework provides one important way of understanding the linkage between cognition and underlying neural systems.

References

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.

Baker, S. C., Rogers, R. D., Owen, A. M., Frith, C. D., Dolan, R. J., Frackowiak, R. S. J., & Robbins, T. W. (1996). Neural systems engaged by planning: A PET study of the Tower of London task. *Neuropsychologia*, *34*, 515.

Beiser, D. G., Hua, S. E., & Houk, J. C. (1997). Network models of the basal ganglia. *Current Opinion in Neurobiology*, *7*, 185.

Bell, A. J., & Sejnowski, T. J. (1997). The independent components of natural images are edge filters. *Vision Research*, *37*, 3327–3338.

Berns, G. S., & Sejnowski, T. J. (1996). How the basal ganglia make decisions. In A. Damasio, H. Damasio, & Y. Christen (Eds.), *Neurobiology of decision-making* (pp. 101–113). Berlin: Springer-Verlag.

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.

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.

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.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual route and parallel-distributed-processing approaches. *Psychological Review*, *100*, 589–608.

Coltheart, M., & Rastle, K. (1994). Serial processing in reading aloud: Evidence for dual-route models of reading. *Journal of Experimental Psychology: Human Perception and Performance*, *20*, 1197–11211.

Contreras-Vidal, J. L., Grossberg, S., & Bullock, D. (1997). A neural model of cerebellar learning for arm movement control: Cortico-spino-cerebellar dynamics. *Learning and Memory*, *3*, 475–502.

Coslett, H. B., & Saffran, E. (1991). Simultanagnosia. To see but not two see. *Brain*, *114*, 1523–1545.

Dayan, P. (1992). The convergenece of TD(λ) for general λ. *Machine Learning*, *8*, 341.

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.

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.

Fukushima, K. (1988). Neocognitron: A hierarchical neural network capable of visual pattern recognition. *Neural Networks*, *1*, 119–130.

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.

Gilbert, C. D. (1996). Plasticity in visual perception and physiology. *Current Opinion in Neurobiology*, *6*, 269.

Goel, V., & Grafman, J. (1995). Are the frontal lobes implicated in ”planning” functions? interpreting data from the tower of hanoi. *Neuropsychologia*, *33*, 623.

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.

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.

Houk, J. C., Davis, J. L., & Beiser, D. G. (Eds.). (1995). *Models of information processing in the basal ganglia*. Cambridge, MA: MIT Press.

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.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, *1*, 541–551.

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.

Livingstone, M., & Hubel, D. (1988). Segregation of form, color, movement, and depth: Anatomy, physiology, and perception. *Science*, *240*, 740–749.

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

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.

Miller, E. K., Erickson, C. A., & Desimone, R. (1996). Neural mechanisms of visual working memory in prefontal 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.

Moll, M., & Miikkulainen, R. (1997). Convergence-zone episodic memory: Analysis and simulations. *Neural Networks*, *10*, 1017–1036.

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

Mozer, M. C. (1991). *The perception of multiple objects: A connectionist approach*. Cambridge, MA: MIT Press.

Mozer, M. C., & Sitton, M. (1998). Computational modeling of spatial attention. In H. Pashler (Ed.), *Attention* (pp. 341–393). London: UCL Press.

Munakata, Y. (1998). Infant perseveration and implications for object permanence theories: A PDP model of the $A\overline{B}$ task. *Developmental Science*, *1*, 161–184.

Munakata, Y., & Stedron, J. M. (in press). Memory for hidden objects in early infancy. In J. Fagen, & H. Hayne (Eds.), *Advances in infancy research, volume 14*. Norwood, NJ: Ablex Publishing Corporation.

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. (1996). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation*, *8*(5), 895–938.

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. (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.* (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., & 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. (2001). Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psychological Review*, *108*, 311–345.

Pinker, S. (1991). Rules of language. *Science*, *253*, 530–535.

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.

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

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.

Pouget, A., & Sejnowski, T. J. (1997). Spatial transformations in the parietal cortex using basis functions. *Journal of Cognitive Neuroscience*, *9*, 222.

Roberts, R. J., & Pennington, B. F. (1996). An interactive framework for examining prefrontal cognitive processes. *Developmental Neuropsychology*, *12*(1), 105–126.

Samsonovich, A., & McNaughton, B. L. (1997). Path integration and cognitive mapping in a continuous attractor neural network model. *Journal of Neuroscience*, *17*, 5900–5920.

Schultz, W., Apicella, P., & Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of Neuroscience*, *13*, 900–913.

Schultz, W., Apicella, P., Romo, R., & Scarnati, E. (1995). Context-dependent activity in primate striatum reflecting past and future behavioral events. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 11–28). Cambridge, MA: MIT Press.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, *275*, 1593.

Schweighofer, N., Arbib, M., & Kawato, M. (1998a). Role of the cerebellum in reaching quickly and accurately: I. A functional anatomical model of dynamics control. *European Journal of Neuroscience*, *10*, 86–94.

Schweighofer, N., Arbib, M., & Kawato, M. (1998b). Role of the cerebellum in reaching quickly and accurately: II. A detailed model of the intermediate cerebellum. *European Journal of Neuroscience*, *10*, 95–105.

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry*, *20*, 11–21.

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

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.

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.

Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review*, *99*, 195–231.

Sutton, R. S. (1988). Learning to predict by the method of temporal diferences. *Machine Learning*, *3*, 9–44.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction.* Cambridge, MA: MIT Press.

Swindale, N. V. (1996). The development of topography in the visual cortex: A review of models. *Network: Computation in Neural Systems*, *7*, 161–247.

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.

Vecera, S. P., & O'Reilly, R. C. (1998). Figure-ground organization and object recognition processes: An interactive account. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 441–462.

Verfaellie, M., Rapcsak, S. Z., & Heilman, K. M. (1990). Impaired shifting of attention in Balint's syndrome. *Brain and Cognition*, *12*, 195–204.

Wickens, J. (1997). Basal ganglia: Structure and computations. *Network: Computation in Neural Systems*, *8*, R77–R109.

Zipser, D., Kehoe, B., Littlewort, G., & Fuster, J. (1993). A spiking network model of short-term active memory. *Journal of Neuroscience*, *13*, 3406–3420.

Glossary

Neural network models: These are models that simulate neurons, typically using simplified approximations (units) that capture the integration of synaptic inputs via weights, and production of an activation output that represents something like the rate of neural firing. Learning mechanisms adapt the synaptic weights to modify the network s function.

Hebbian learning: A learning mechanism that increases the weights between neural units in proportion to the sending and receiving activation (the units that fire together wire together). Provisions for decreasing weights for units that do not fire together are typically used as well. This is typically unsupervised learning, in that it works directly from the inputs without target patterns (see error-driven learning).

Error-driven learning: Learning that adapts weights to minimize overall error as recorded on output layers of the network. The error is often defined by comparing outputs to specified target patterns (called supervised learning), but other

Inhibitory competition: Where neurons inhibit each other and therefore compete for activation — the most strongly activated neurons will inhibit more weakly activated ones.

Interactive/recurrent activations: This is where two sets of neural activations interact by providing mutual input to each other (e.g., unit A activates unit B, and B activates A in turn).

Receptive field: The set of stimuli that activate a neuron, as determined by its pattern of synaptic connections or by recording its actual responses.

Word processing package

Originally written in LaTeX under Linux, translated into RTF and then into Word under Office 2000 on a PC (Windows ’98).