Modeling Integration and Dissociation in Brain and Cognitive Development

Randall C. O'Reilly
Department of Psychology
University of Colorado Boulder
345 UCB
Boulder, CO 80309
oreilly@psych.colorado.edu
word count: 8,600 main text + 1,113 refs + 1000 figs = 10,713 total

Abstract:

Over the course of development, brain areas can become increasingly dissociated in their functions, or increasingly integrated. Computational models can provide insights into how and why these opposing effects happen. This paper presents a computational framework for understanding the specialization of brain functions across the hippocampus, neocortex, and basal ganglia. This framework is based on computational tradeoffs that arise in neural network models, where achieving one type of learning function requires very different parameters from those necessary to achieve another form of learning. For example, we dissociate the hippocampus from cortex with respect to general levels of activity, learning rate, and level of overlap between activation patterns. Similarly, the frontal cortex and associated basal ganglia system have important neural specializations not required of the posterior cortex system. Taken together, these brain areas form an overall cognitive architecture, which has been implemented in functioning computational models, provides a rich and often subtle means of explaining a wide range of behavioral and cognitive neuroscience data. The developmental implications of this framework, and other computational mechanisms of dissociation and integration, are reviewed.

Introduction

The brain is not a homogeneous organ: different brain areas are specialized for different cognitive functions. On the other hand, it is also clear that the brain does not consist of strictly encapsulated modules with perfectly segregated contents. This paper reviews one approach to understanding the nature of specialized functions in terms of the logic of *computational tradeoffs* in neural network models of brain areas. The core idea behind this approach is that different brain areas are specialized to satisfy fundamental tradeoffs in neural network's performance of different kinds of learning and memory tasks. This way of characterizing the specializations of brain areas is generally consistent with some other theoretical frameworks, but it offers a level of precision and subtlety suitable for understanding complex interactions between different brain areas.

Countering these specialization pressures is the need to integrate information to avoid the well-known *binding problem* that arises with completely segregated representations. For example, if color and shape information are encoded by distinct neural populations, it then becomes difficult to determine which color goes with which shape when multiple objects are simultaneously present in the stimulus input. One popular solution to this problem is to invoke the mechanism of synchronous neural firing, such that stimulus features corresponding to the same object fire together, and out of phase with those for other objects (e.g., von der Malsburg, 1981; Gray, Engel, Konig, & Singer, 1992; Engel, Konig, Kreiter, Schillen, & Singer,

in press in: Y. Munakata & M.H. Johnson (Eds) *Processes of Change in Brain and Cognitive Development: Attention and Performance XXI*. Oxford University Press. Supported by ONR grant N00014-03-1-0428, and NIH grants MH069597 and MH64445.

1992; Zemel, Williams, & Mozer, 1995; Hummel & Biederman, 1992). However, there are a number of problems with this approach, as elaborated below. One alternative is to use conjunctive representations, where individual neural representations encode multiple stimulus features (e.g., one unit might encode the conjunction of "blue" and "triangle"). This solution, in its simple form, is also highly problematic, producing a combinatorial explosion of different representations for each possible conjunction, and the inability to generalize knowledge across different experiences. There is a more subtle and powerful form of conjunctive representations, however, known as distributed coarse-coded conjunctive representations, which avoid these problems (Hinton, McClelland, & Rumelhart, 1986; Wickelgren, 1969; Seidenberg & McClelland, 1989; St John & McClelland, 1990; Mozer, 1991; Mel & Fiser, 2000; O'Reilly & Soto, 2002; O'Reilly, Busby, & Soto, 2003). Individual units in such representations encode multiple subsets of conjunctions (i.e., coarse-coding), and the distributed pattern of activation across many such units serves to distinguish different stimulus configurations. This type of representation is ubiquitous in the brain, and its computational features are explored later in this paper.

Taking these two forces of integration and dissociation together, a clear reconciliation emerges. Instead of viewing brain areas as being specialized for specific *representational content* (e.g., color, shape, location, etc), areas are specialized for specific *computational functions* by virtue of having different neural parameters. Within each area, many types of representational content are intermixed in distributed coarse-coded conjunctive representations, to avoid the binding problem. This framework flies in the face of the pervasive tendency to associate brain areas with content (e.g., the fusiform face area (Kanwisher, 2000); the ventral what pathway vs. the dorsal where pathway (Ungerleider & Mishkin, 1982); the hippocampus as a spatial map (O'Keefe & Nadel, 1978), etc). Instead it is aligned with alternative frameworks that focus on function. For example, the dorsal "where" pathway has been reinterpreted as "vision for action", which integrates both "what" and "where" information in the service of performing visually-guided motor actions (Goodale & Milner, 1992). Similarly, the fusiform face area has been characterized instead as an area suitable for subordinate category representations of large numbers of similar items, which includes faces but also birds in the case of bird experts, for example (Tarr & Gauthier, 2000). Below, the case for understanding the hippocampus as a system specialized for the general function of rapid learning of arbitrary conjunctive information, including but not restricted to spatial information, is reviewed (O'Reilly & McClelland, 1994; McClelland, McNaughton, & O'Reilly, 1995; O'Reilly & Rudy, 2001; Norman & O'Reilly, 2003).

This "functionalist" perspective has been instantiated in a number of neural network models of different brain areas, including posterior (perceptual) neocortex, hippocampus, and the prefrontal cortex/basal ganglia system. We are now in the process of integrating these different models into an overall biologically-based cognitive architecture (Figure 1). Each component of the architecture is specialized for a different function by virtue of having different parameters and neural specializations (as motivated by computational tradeoffs), but the fundamental underlying mechanisms are the same across all areas. Specifically, our models are all implemented within the Leabra framework (O'Reilly, 1998; O'Reilly & Munakata, 2000), which includes a coherent set of basic neural processing and learning mechanisms that have been developed by different researchers over the years. Thus, many aspects of these areas work in the same way (and on the same representational content), and in many respects the system can be considered to function as one big undifferentiated whole. For example, any given memory is encoded in synapses distributed throughout the entire system, and all areas participate in some way in representing most memories. Therefore, this architecture is much less modular than most conceptions of the brain, while still providing a principled and specific way of understanding the differential contributions of different brain areas. These seemingly contradictory statements are resolved through the process of developing and testing concrete computational simulations that help us understand the ways in which these areas contribute differentially, and similarly, to cognitive and behavioral functions.

In the remainder of the paper, the central computational tradeoffs underlying our cognitive architecture

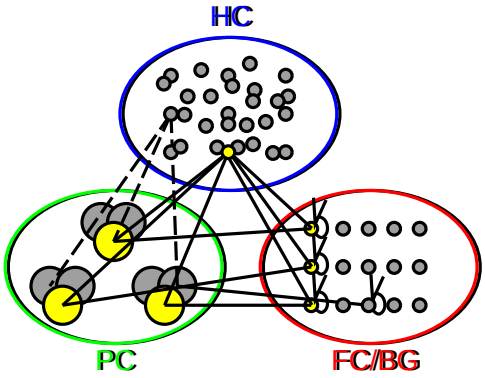


Figure 1: Tripartite cognitive architecture defined in terms of different computational tradeoffs associated with Posterior Cortex (PC), Hippocampus (HC) and Frontal Cortex (FC) (with motor frontal cortex constituting a blend between FC and PC specializations). Large overlapping circles in PC represent overlapping distributed representations used to encode semantic and perceptual information. Small separated circles in HC represent sparse, pattern-separated representations used to rapidly encode ("bind") entire patterns of information across cortex while minimizing interference. Isolated, self-connected representations in FC represent isolated stripes (columns) of neurons capable of sustained firing (i.e., active maintenance or working memory). The basal ganglia also play a critical role in the FC system by modulating ("gating") activations there based on learned reinforcement history.

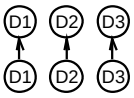
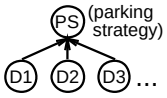
are reviewed, along with a more detailed discussion of the binding problem and the distributed coarse-coded representations solution to it. In each case, these ideas are applied to relevant developmental phenomena, where they may have some important implications, despite the fact that these ideas have been largely based on considerations from the adult system (though across multiple species). There are also some important computational mechanisms of integration and dissociation that do not emerge directly from this computational tradeoff framework, which are briefly reviewed.

Specializations in Hippocampus and Posterior Neocortex

One of the central tradeoffs behind our approach involves the process of learning novel information rapidly without interfering catastrophically with prior knowledge. This form of learning requires a neural network with very sparse levels of overall activity (leading to highly separated representations), and a relatively high learning rate (i.e., high levels of synaptic plasticity). These features are incompatible with the kind of network that is required to acquire general statistical information about the environment, which needs highly overlapping, distributed representations with relatively higher levels of activity, and a slow rate of learning. The conclusion we have drawn from this mutual incompatibility (see Figure 2a for a summary) is that the brain must have two different learning systems to perform these different functions (O'Reilly & McClelland, 1994; McClelland et al., 1995; O'Reilly & Rudy, 2001; Norman & O'Reilly, 2003). This computational tradeoff idea fits quite well with a wide range of existing theoretical ideas and converging cognitive neuroscience data on the properties of the hippocampus and posterior neocortex, respectively (Scoville & Milner, 1957; Marr, 1971; Grossberg, 1976; O'Keefe & Nadel, 1978; Teyler & Discenna, 1986; McNaughton & Morris, 1987; Sherry & Schacter, 1987; Rolls, 1989; Sutherland & Rudy, 1989; Squire, 1992; Eichenbaum, Otto, & Cohen, 1994; Treves & Rolls, 1994; Burgess & O'Keefe, 1996; Wu, Baxter, & Levy, 1996; Moll & Miikkulainen, 1997; Hasselmo & Wyble, 1997; Aggleton & Brown, 1999; Yonelinas, 2002).

We have instantiated our theory in the form of a computational model of the hippocampus and neocortex (Figure 2b). This same model has been extensively tested through applications to a wide range of data from

a)

	Two Incompatible Goals	
	Remember Specifics	Extract Generalities
Example:	Where is car parked?	Best parking strategy?
Need to:	Avoid interference	Accumulate experience
	Solution:	
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)
	These are incompatible, need two different systems:	
System:	Hippocampus	Neocortex

b)

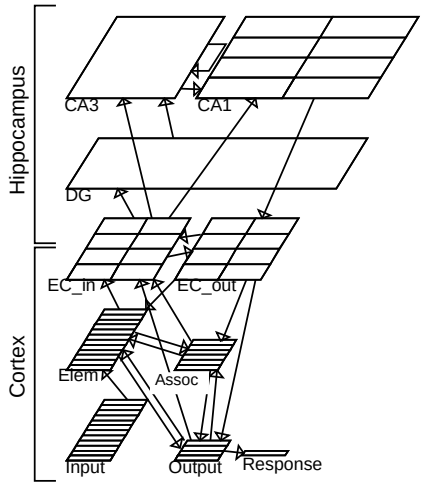


Figure 2: **a)** Computational motivation for two complementary learning & memory systems in the brain: There are two incompatible goals that such systems need to solve. One goal is to remember specific information (e.g., where one's car is parked). The other is to extract generalities across many experiences (e.g., developing the best parking strategy over a number of different days). The neural solutions to these goals are incompatible: Memorizing specifics requires separate representations that are learned quickly, and automatically, while extracting generalities 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. **b)** Our hippocampal/cortical model (O'Reilly & Rudy, 2001, Norman & O'Reilly, 2003). The cortical system consists of sensory input pathways (including elemental (Elem) sensory coding and higher-level association cortex (Assoc)) and motor output. These feed via the entorhinal cortex (EC_in, superficial layers of EC) into the hippocampus proper (dentate gyrus (DG), the fields of Ammon's horn (CA3, CA1), which in turn project back to cortex via EC_out (deep layers of EC). The DG and CA3 areas have particularly sparse representations (few neurons active), which enables rapid learning of arbitrary conjunctive information (i.e., "episodic learning") by producing pattern separation and thus minimizing interference.

humans and animals (O'Reilly, Norman, & McClelland, 1998; O'Reilly & Rudy, 2001; Norman & O'Reilly, 2003; Rudy & O'Reilly, 2001; Frank, Rudy, & O'Reilly, 2003) (see O'Reilly & Norman, 2002 for a concise review). The hippocampal model performs encoding and retrieval of memories in the following manner: During encoding, the hippocampus develops relatively non-overlapping (pattern-separated) representations of cortical inputs (communicated via entorhinal cortex, EC) in region CA3 (strongly facilitated by the very sparse dentate gyrus (DG) inputs). Active units in CA3 are linked to one another (via Hebbian learning), and to a sparser but stable re-representation of the EC input pattern in region CA1. During retrieval, presentation of a partial version of a previously encoded memory representation leads to reconstruction of the complete original CA3 representation (supported by Hebbian-strengthened connections within CA3, and other synaptic modifications throughout the hippocampus). This is *pattern completion*, which is essentially cued recall, where an entire representation is completed or filled-in based on a partial cue. As a consequence of this pattern completion in CA3, the entire studied pattern on the EC output layer is reconstructed (via area CA1), which then spreads out to cortex to fully represent the recalled information. As reviewed in Norman and O'Reilly (2003) and O'Reilly and Rudy (2001), our hippocampal model closely resembles other neural network models of the hippocampus (Treves & Rolls, 1994; Touretzky & Redish, 1996; Burgess & O'Keefe, 1996; Wu et al., 1996; Moll & Miikkulainen, 1997; Hasselmo & Wyble, 1997). There are differences, but the family resemblance between these models far outweighs the differences. Recent data comparing neural activation patterns in CA3 and CA1 clearly supports the model's distinctions between these two areas, where

CA3 is subject to more pattern completion and separation, while CA1 is a more stable but sparser encoding of the current inputs (Lee, Yoganarasimha, Rao, & Knierim, 2004; Vazdarjanova & Guzowski, in press).

In contrast with the rapid, conjunctive learning supported by the hippocampus, our cortical model can support generalization across a large number of experiences, as a result of two neural properties. First, our simulated cortical neurons have a slow learning rate (i.e., small changes in synaptic efficacy after a single presentation of a stimulus). That property insures that any single event has a limited effect on cortical representations. It is the gradual accumulation of many of these small impacts that shapes the representation to capture things that are reliably present across many experiences (i.e., the general statistical structure or regularities of the environment). Second, our model employs representations that involve a relatively large number of neurons (e.g., roughly 15-25%). This property increases the probability that similar events will activate overlapping groups of neurons, thereby enabling these neurons to represent the commonalities across many experiences. More discussion of cortical learning and development is presented later.

Hippocampal and Cortical Contributions to Recall and Recognition Memory

To flesh out some of the implications of this approach, we briefly review the application of this model to human memory, where we can understand the distinction between recall and recognition memory (Norman & O'Reilly, 2003). The key result is that the ability of the hippocampus to rapidly encode novel conjunctive information with minimal interference is critical for supporting recall of detailed information from prior study episodes. In contrast, the cortex, even with a slow learning rate, can contribute to the recognition of previously experienced stimuli by providing a global, scalar *familiarity* signal. This familiarity-based recognition does not require the ability to pattern-complete missing elements of the original study episode. Instead, it simply requires some kind of ability to match the current input with an existing representation, and report something akin to the "global-match" between them (e.g., Hintzman, 1988; Gillund & Shiffrin, 1984). It turns out that our cortical network can support this recognition function as a result of small "tweaks" to the weights of existing representations in the network. These small weight changes cause a recently-activated cortical representation to be somewhat "sharper" than before (i.e., the difference between active and inactive units is stronger; the contrast is enhanced). This difference in sharpness can be reliably used to distinguish "old" from "new" items in recognition memory tests.

This distinction between hippocampal recall and cortical recognition is consistent with many converging sources of data, as reviewed in Yonelinas (2002). One of the interesting novel predictions that arose from our model is that input stimulus similarity and recognition test format should critically impact the cortical system, but not the hippocampal system. Specifically, as the similarity of input stimuli increases, the corresponding cortical representations will also increase in overlap, and this will cause the cortical recognition signal (sharpness) to also overlap. Thus, on a recognition memory test using novel test stimuli that overlap considerably with studied items (e.g., study "CAT" and test with "CATS"), the cortical system would be much more likely to false alarm to these similar lures. In contrast, the pattern separation property of the hippocampal system will largely prevent this similarity-based confusion, by encoding the patterns with relatively less overlapping internal representations. However, if both the studied item and the similar lure were presented together at test in a forced-choice testing paradigm, then the cortical system can still provide good performance. This is because although the similar lure will activate an overlapping cortical representation, this representation will nevertheless be reliably less sharpened than that of the actual studied item.

These predictions from the computational model have been tested in experiments on a patient (YR) with selective hippocampal damage, and matched controls (Holdstock, Mayes, Roberts, Cezayirli, Isaac, O'Reilly, & Norman, 2002). YR is a 61-year-old woman that had focal hippocampal damage due to a painkiller overdose. The damage did not extend to the surrounding medial temporal lobe cortex. On the yes/no recognition task, images were presented one at a time, and the subjects had to respond "yes" if the image was seen in the previous study phase. On the forced-choice recognition task, a studied image was

presented with two novel ones, and the subjects were asked to find the studied one. YR was impaired relative to controls only on the yes/no recognition test with similar lures, and not on the forced-choice test with similar lures, or either test with dissimilar lures. Furthermore, she was impaired at a recall test matched for difficulty with the recognition tests in the control group. This pattern matches exactly the predictions of the model with respect to the impact of a selective hippocampal lesion.

There are numerous other examples where the predictions from our computational models have been tested in both humans and animals (O'Reilly et al., 1998; O'Reilly & Rudy, 2001; Norman & O'Reilly, 2003; Rudy & O'Reilly, 2001; Frank et al., 2003). In many ways, the understanding we have achieved through these computational models accords well with theories derived through other motivations. For example, there is broad agreement among theorists that a primary function of the hippocampus is the encoding of episodic or spatial memories (e.g., Vargha-Khadem, Gadian, Watkins, Connelly, Van Paesschen, & Mishkin, 1997; Squire, 1992; O'Keefe & Nadel, 1978). This function emerges from the use of sparse representations in our models, because these representations cause the system to develop conjunctive representations that bind together the many different features of an episode or location into a unitary encoding (e.g., O'Reilly & Rudy, 2001; O'Reilly & McClelland, 1994). However, the models are also often at variance with existing theorizing. For example, the traditional notions of "familiarity" and "recall" do not capture all the distinction between neocortical and hippocampal contributions, as we showed in a number of cases in Norman and O'Reilly (2003). For example, neocortical representations can be sensitive to contextual information, and even to arbitrary paired associates, which is not well accounted for by traditional notions of how the familiarity system works.

Developmental Implications

Some implications of this overall framework for understanding various developmental phenomena were described by Munakata (2004). One intriguing application is to the phenomenon of infantile amnesia, where most people cannot remember any experiences prior to the age of about 2-3 years (Howe & Courage, 1993). As with many accounts of this phenomenon, she argues that representational change in the cortex during this formative period can result in the inability to retrieve hippocampal episodic representations later in life (e.g., McClelland et al., 1995). However, this general account does not explain why it is that this representational change does not render all forms of knowledge inaccessible; why does it seem to specifically affect hippocampal episodic memories? Munakata (2004) argues that the pattern separation property of the hippocampus makes it especially sensitive to even relatively small changes in cortical representations. By contrast, the cortex itself would be much less sensitive to such changes, because it tends to generalize across similar patterns to a much greater extent.

Another potential application of this framework is in the domain of so-called "fast-mapping" phenomena, where children are capable of rapid (e.g., one-trial) learning of novel information (Hayne, this volume; Hayne, Boniface, & Barr, 2000; Markson, this volume; Bloom & Markson, 1998). In the case of the mobile-conjugate reinforcement learning and deferred imitation studies of Hayne and colleagues, infants and children exhibit one-trial learning that is highly sensitive to the study/test stimulus overlap, for both task-relevant and irrelevant stimulus features. This sensitivity to pattern overlap (and fast learning) is highly suggestive of hippocampal function, where the sparse activity levels result in units that are sensitive to stimulus conjunctions (O'Reilly & Rudy, 2001) — only if the study and test environments have sufficient similarity will pattern completion be triggered to produce successful recall. Otherwise, pattern separation will result in an inability to recall the study episode. Nevertheless, there is some question as to when the hippocampus becomes functional in human development, and it is also possible that the high degree of plasticity in the infant neocortex could support rapid learning of this sort. However, the apparently highly conjunctive nature of this fast learning, which fits so well with the hippocampal mechanisms, remains to be explained under this account. Computational models of the detailed behavioral results would be useful to

explore these alternative hypotheses.

The fast mapping phenomena studied by Markson and colleagues in the context of early word learning may reflect a more complex interaction between cortical and hippocampal learning mechanisms. This is because this form of learning appears to support considerable generalization and inference, which are hallmarks of cortical representations. Thus, the hippocampus in this case may be only responsible for linking a word with otherwise fairly well-developed cortical representations of the underlying perceptual world. As we saw in the case of recognition memory, the cortical system can exhibit behaviorally-measurable one-trial learning, as long as this learning involves small changes to largely existing representations. Therefore, word-learning fast mapping may be best explained as relatively small changes in the landscape of existing semantic representations, which serve to bring some latent representations "over threshold", while the hippocampus helps in the linking of these semantic representations with an associated arbitrary verbal label. Again, this is a rich domain that is just waiting to be explored from this hippocampus/cortex computational modeling framework.

The Prefrontal Cortex/Basal Ganglia System

The same tradeoff logic applied to the hippocampal/cortical system has been applied to understanding the specialized properties of the frontal cortex (particularly focused on the prefrontal cortex, PFC) relative to the posterior neocortex and hippocampal systems. The tradeoff in this case involves specializations required for maintaining information in an active state (i.e., maintained neural firing) relative to those required for performing semantic associations and other forms of inferential reasoning. Specifically, active maintenance (often referred to by the more general term of working memory) requires relatively isolated representations so that information does not spread out and get lost over time (O'Reilly & Munakata, 2000; O'Reilly, Braver, & Cohen, 1999). In contrast, the overlapping distributed representations of posterior cortex support spreading associations and inference by allowing one representation to activate aspects of other related representations (e.g., McClelland & Rogers, 2003; Lambon-Ralph, Patterson, Garrard, & Hodges, 2003). This tradeoff is illustrated and described further in Figure 3. Neural anatomy and physiology data from prefrontal cortex in monkeys is consistent with this idea. Specifically, prefrontal cortex has relatively isolated "stripes" of interconnected neurons (Levitt, Lewis, Yoshioka, & Lund, 1993), and neurons located close by each other all maintain the same information according to electrophysiological recordings of "iso-coding microcolumns" (Rao, Williams, & Goldman-Rakic, 1999).

In addition to relatively isolated patterns of connectivity, the prefrontal cortex may be specialized relative to posterior cortex by virtue of its need for an adaptive gating mechanism. This mechanism dynamically switches between rapidly updating new information (gate open) and robustly maintaining other information (gate closed) (Figure 4a). (Cohen, Braver, & O'Reilly, 1996; Braver & Cohen, 2000; O'Reilly et al., 1999; O'Reilly & Munakata, 2000). This adaptive gating also needs to be selective, such that some information is updated while other information is maintained. This can be achieved through the parallel loops of connectivity through different areas of the basal ganglia and frontal cortex (Figure 4b) (Alexander, DeLong, & Strick, 1986; Graybiel & Kimura, 1995; Middleton & Strick, 2000). We postulate that these parallel loops also operate at the finer level of the isolated anatomical stripes in prefrontal cortex, and provide a mechanism for selectively updating the information maintained in one stripe, while robustly maintaining information in other stripes.

A detailed computational model of how such a system would work, and how it can learn which stripes to update when, has been developed (O'Reilly & Frank, submitted). This model avoids the "homunculus problem" that arises in many theories of prefrontal cortex, where it is ascribed powerful "executive functions" (e.g., Baddeley, 1986) that remain mechanistically unspecified. In effect, these theories rely on unexplained human-like intelligence in the PFC, amounting to a "homunculus" (i.e., a small man inside the head). In

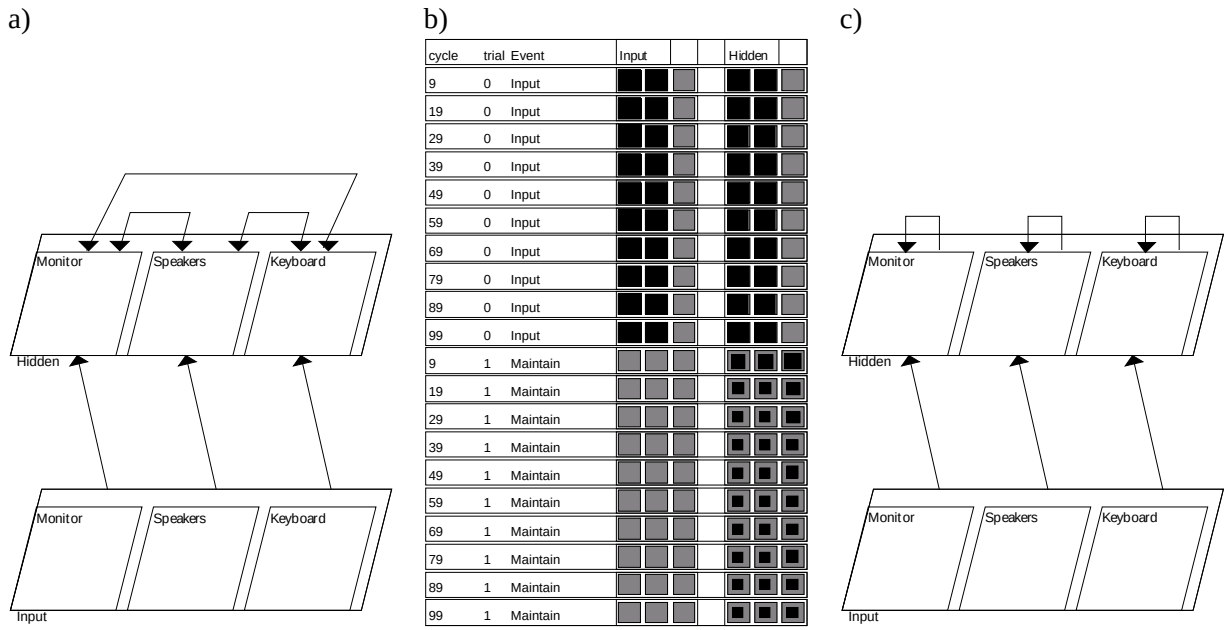


Figure 3: Demonstration of the tradeoff between interconnected and isolated neural connectivity and inference vs. active maintenance. **a)** Interconnected network: Weights (arrows) connect hidden units that represent semantically related information. Such connectivity could subserve semantic networks of posterior cortical areas. **b)** Input and hidden unit activity as the interconnected network is presented with two inputs (top half of figure) and then those inputs are removed (bottom half of figure). Each row corresponds to one time step of processing. Each unit's activity level is represented by the size of the corresponding black square. The network correctly activates the corresponding hidden units when the inputs are present, but fails to maintain this information alone when the input is removed, due to interactive representations. **c)** Network with isolated representations: Each hidden unit connects to only itself, rather than to other semantically-related units, and thus information does not spread over time, supporting robust active maintenance abilities associated with prefrontal cortical areas. Adapted from O'Reilly & Munakata (2000).

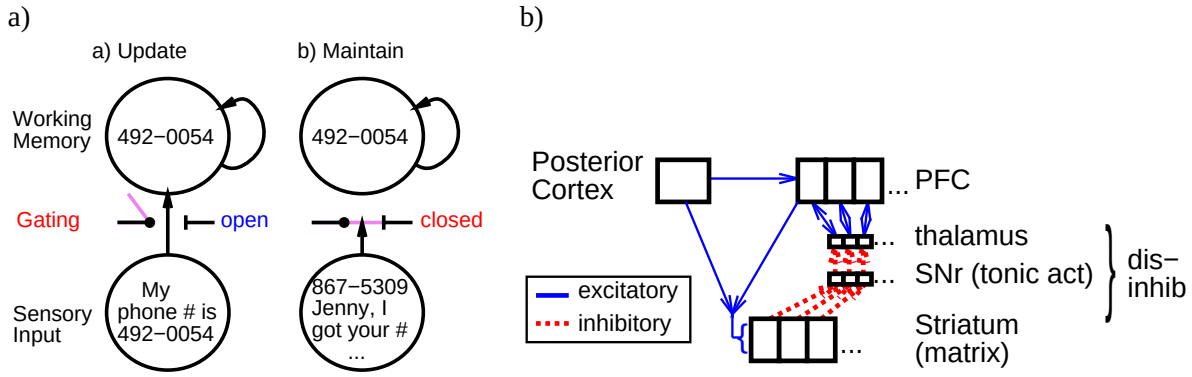


Figure 4: **a)** Illustration of adaptive gating. When the gate is open, sensory input can rapidly update working memory (e.g., encoding the cue item A in the 1-2-AX task), but when it is closed, it cannot, thereby preventing other distracting information (e.g., distractor C) from interfering with the maintenance of previously stored information. **b)** The basal ganglia (striatum, globus pallidus and thalamus) are interconnected with frontal cortex through a series of parallel loops. Striatal neurons disinhibit prefrontal cortex by inhibiting tonically active substantia nigra pars reticulata (SNr) neurons, releasing thalamic neurons from inhibition. This disinhibition provides a modulatory or gating-like function.

contrast, our model learns to solve complex working memory tasks starting with no preexisting knowledge whatsoever, demonstrating that they are capable of developing powerful forms of intelligence autonomously.

Development of Rule-Like PFC Representations

We have begun to explore some of the developmental implications of the above specialized PFC/BG mechanisms. In particular, the presence of a adaptive gating mechanism can impose important constraints on the types of representations that form in the PFC system, which in turn can impact the overall behavior of the system in important ways. We recently showed that a network having an adaptive gating mechanism developed abstract, rule-like representations in its simulated PFC, whereas models lacking this mechanism did not (Rougier, Noelle, Braver, Cohen, & O'Reilly, submitted). Furthermore, the presence of these rule-like representations resulted in greater flexibility of cognitive control, as measured by the ability to generalize knowledge learned in one task context to other tasks. As elaborated below, these results may have important implications for understanding the nature of development in the PFC, and how it can contribute to tasks in ways that are not obviously related to working memory function (e.g., by supporting more regular, rule-like behavior).

Rougier et al. (submitted) trained a range of different models on a varying number of related tasks operating on simple visual stimuli (e.g., *name* a "feature" of the stimulus along a given "dimension" such as its color, shape, or size; *match* two stimuli along one of these dimensions; *compare* the relative size of two stimuli). Though simple, these tasks also allowed us to simulate benchmark tasks of cognitive control such as Wisconsin card sorting (WCST) and the Stroop task. The generalization test for the cognitive flexibility of the models involved training a given task on a small percentage (e.g., 30%) of all the stimuli, and then testing that task on stimuli that were trained in other tasks. To explore the impact of the adaptive gating mechanism and other architectural features, a range of models having varying numbers of these features were tested.

The model with the full set of prefrontal working memory mechanisms (including adaptive gating) achieved significantly higher levels of generalization than otherwise comparable models that lacked these specialized mechanisms. Furthermore, this benefit of the prefrontal mechanisms interacted with the breadth of experience the network had across a range of different tasks. The network trained on all four tasks generalized significantly better than one trained on only pairs of tasks, but this was only true for the full PFC model. These results were strongly correlated ($r = .97$) with the extent to which the model developed abstract rule-like representations of the stimulus dimensions that were relevant for task performance. Thus, the model exhibited an interesting interaction between nature (the specialized prefrontal mechanisms) and nurture (the breadth of experience): both were required to achieve high levels of generalization.

There are numerous points of contact between this model and a range of developmental and neuroscience data. For example, the need for extensive breadth of experience in the model to develop more flexible cognitive function may explain the why the prefrontal cortex requires such an extended period of development (up through late adolescence; Casey, Durston, & Fossella, 2001; Morton & Munakata, 2002b; Lewis, 1997; Huttenlocher, 1990). That is, the breadth of experience during that time enables the PFC to develop systematic representations that support the flexible reasoning abilities we have as adults. This model is also consistent with data showing that damage to prefrontal cortex impairs abstraction abilities (e.g., Dominey & Georgieff, 1997), and that prefrontal cortex in monkeys develops more abstract category representations than those in posterior cortex (Wallis, Anderson, & Miller, 2001; Freedman, Riesenhuber, Poggio, & Miller, 2002; Nieder, Freedman, & Miller, 2002). Furthermore, the growing literature on developing task switching abilities in children should prove to be a useful domain in which to explore the developmental properties of this model (e.g., Zelazo, Frye, & Rapus, 1996; Munakata & Yerys, 2001; Morton & Munakata, 2002a, 2002b).

In our current research with this PFC/BG model, we are expanding the range and complexity of cognitive

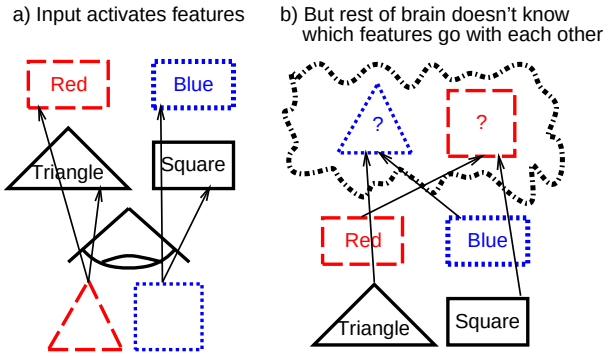


Figure 5: Illustration of the binding problem. a) Visual inputs (red triangle, blue square) activate separate representations of color and shape properties. b) However, just the mere activation of these features does not distinguish for the rest of the brain the alternative scenario of a blue triangle and a red square. Red is indicated by dashed outline and blue by a dotted outline.

tasks, and in the process undertaking an exploration of the "educational curriculum" that we present to the model. Specifically, we are trying to build up to a wide range of tasks through the training of a smaller set of core competencies. We are starting with a simple sensory/motor domain where the tasks involve focusing on subsets of the visual inputs, and producing appropriate verbal and/or motor outputs. For example, the network is being trained to name, match, point, etc. according to different stimulus dimensions or locations. We plan to take this process one step further in the course of developing the full tripartite cognitive architecture, which will involve a more sophisticated perceptual system capable of operating on raw bitmap images, to perform more complex tasks such as visual search in cluttered environments, and real-world navigation. This developmental approach to constructing our models is a necessary consequence of the fact that they are fundamentally learning models. They start out with only broad parametric preconfiguration, and then must develop their sophisticated abilities through experience-driven learning. Thus, these models should provide an interesting test-bed for understanding how such parametric variations across different areas of the network lead to differentiations in mature function (e.g., Elman, Bates, Johnson, Karmiloff-Smith, Parisi, & Plunkett, 1996).

The Need for Integration: Binding

To this point, we have focused on the ways in which neural systems need to be specialized to carry out different computational functions. However, there are opposing pressures that force the integration of information processing functions within a single brain area. In particular, as noted earlier, the binding problem places important demands on how information is represented within a given brain area, requiring information to be integrated. As shown in Figure 5, the binding problem arises whenever different aspects of a stimulus input (e.g., color and shape) are encoded by separate neural units. When you have two or more inputs, then you cannot recover from the internal representation which color goes with which shape: was it a red triangle or a blue triangle out there in the world? Although the discussion below focuses on the domain of posterior cortical sensory representations, these binding issues are important for virtually all brain areas.

One trivial solution to the binding problem is to use conjunctive representations to represent each binding that the system needs to perform. In the example shown in Figure 5, there would be a particular unit that codes for a blue square and another that codes for a red triangle. While it is intuitively easy to understand how such conjunctive representations solve the binding problem, they are intractable because they produce a combinatorial explosion in the number of units required to code for all possible bindings as the number

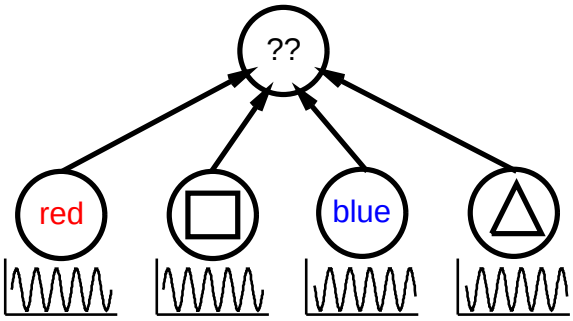


Figure 6: Decoding problem for temporal synchrony. Two sets of features are each fi ring in phase with each other, and out of phase with the other set (as indicated by the sine wave plots below the features). Without additional mechanisms, it is unclear how a downstream neuron can decode this information to determine what is actually present: it is being uniformly driven by synaptic input at all phases, and its activation would be the same for any combination of synchrony in the input features. Also, even though it looks like the synchronous fi ring is discriminable, both sets of units have synchronous fi ring, so there is no basis to choose one over another. One solution is to build in preferential weights for one set of features (e.g., "red square") but this amounts to a conjunctive representation, which the temporal synchrony approach is designed to avoid in the fi rst place.

of features to be bound increases. As an example, assume that all objects in the world can be described by 32 different dimensions (e.g., shape, size, color, etc), each of which contains 16 different feature values. To encode all possible bindings using the naive approach, 16^{32}, or $3.5x10^{38}$ units would be needed. If the system needed to bind features for 4 objects simultaneously, 4 times as many units would be needed. Of course, the brain binds many more types of features and does so with far less units.

Temporal synchrony is a popular alternative to simple conjunctive approach to binding (e.g., von der Malsburg, 1981; Gray et al., 1992; Engel et al., 1992; Zemel et al., 1995; Hummel & Biederman, 1992). This account holds that when populations of neurons that represent various features fire together, those features are considered bound together. To encode multiple distinct sets of bindings, different groups of neurons fire at different phase offsets within an overall cycle of firing, using time to separate the different representations. In the example of Figure 5, the "red" and "triangle" units would fire together, and out of phase with the "blue" and "square" units. This temporal interleaving is appealing in its simplicity, and the many reports of coherent, phasic firing of neurons in the brain appear to lend it some credibility (e.g., Gray et al., 1992; Engel et al., 1992; Csibra, Davis, & Johnson, 2000).

However, the temporal synchrony account has several problems, as detailed in several existing critiques (O'Reilly et al., 2003; Cer & O'Reilly, in press; Shadlen & Movshon, 1999). For example, the transience of temporal synchrony causes problems when bound information needs to be encoded in long-term memory. One proposal is that there is a separate conjunctive representation system for everything that is encoded into long term memory (Hummel & Holyoak, 1997), with the idea that this is a small enough set that the combinatorial explosion of such conjunctions is not a problem. However, there is considerable evidence that just about every activation state in our brains produces a lasting trace in the synaptic connections that can later be measured in priming or perceptual learning studies (e.g., Furmanski & Engel, 2000; Gilbert, Sigman, & Crist, 2001; Adini, Sagi, & Tsodyks, 2002; Aslin, Blake, & Chun, 2002; Wagner, Koutstaal, Maril, Schacter, & Buckner, 2000; Stark & McClelland, 2000)— this would suggest that combinatorial explosion is a problem. Furthermore, the process of actually using ("decoding") the temporal synchrony binding information is problematic as shown in Figure 6. In addition, the data showing synchronous neural firing falls well short of demonstrating the interleaved phase-offset synchrony necessary for binding. Instead, this data may just be an epiphenomenon of spike-based neural firing dynamics.

Fortunately, there is another alternative way of solving the binding problem, which involves a more

obj1	obj2	R	G	B	S	C	T	RC GS BT
RS	GC	1	1	0	1	1	0	0
RC	GS	1	1	0	1	1	0	1
RS	GT	1	1	0	1	0	1	0
RT	GS	1	1	0	1	0	1	1
RS	BC	1	0	1	1	1	0	0
RC	BS	1	0	1	1	1	0	1
RS	BT	1	0	1	1	0	1	1
RT	BS	1	0	1	1	0	1	0
RC	GT	1	1	0	0	1	1	1
RT	GC	1	1	0	0	1	1	0
RC	BT	1	0	1	0	1	1	1
RT	BC	1	0	1	0	1	1	0
GS	BC	0	1	1	1	1	0	1
GC	BS	0	1	1	1	1	0	0
GS	BT	0	1	1	1	0	1	1
GT	BS	0	1	1	1	0	1	0
GC	BT	0	1	1	0	1	1	1
GT	BC	0	1	1	0	1	1	0

Table 1: Solution to the binding problem by using representations that encode combinations of input features (i.e., color and shape), but achieve greater efficiency by representing multiple such combinations. Obj1 and obj2 show the features of the two objects. The first six columns show the responses of a set of representations that encode the separate color and shape features: R = Red, G = Green, B = Blue, S = Square, C = Circle, T = Triangle. Using only these separate features causes the binding problem: observe that the two configurations in each pair are equivalent according to the separate feature representation. The final unit encodes a combination of the three different conjunctions shown at the top of the column, and this is enough to disambiguate the otherwise equivalent representations.

efficient way of implementing conjunctive representations using distributed coarse-coded conjunctive representations (DCC) (Cer & O'Reilly, in press; Mel & Fiser, 2000). A DCC representation encodes binding information via a number of simultaneously active units (i.e., a distributed representation; Hinton et al., 1986), where each unit is activated by multiple different conjunctions. For example, a given unit might respond to red+circle *or* green+square *or* blue+triangle. By getting more conjunctive mileage out of each unit, and leveraging the combinatorial power of distributed representations across multiple units, this solution can be much, much more efficient than naive conjunctive representations (Table 1). For example, for the 32 dimensions with 16 features each case mentioned above, only 512 units would be required under an optimal binary distributed representation (see Cer and O'Reilly (in press) for details). The numbers for more realistic neural networks would certainly be higher than this, but nowhere near the $3.5x10^{38}$ units of the simple conjunctive approach. In addition to this efficiency, virtually every neural recording study ever performed supports these DCC representations, in that individual neurons inevitably encode conjunctions of different stimulus/task features (e.g., Tanaka, 1996; Rao, Rainer, & Miller, 1997; Barone & Joseph, 1989; Ito, Westheimer, & Gilbert, 1998; Walker, Ohzawa, & Freeman, 1999).

Spatial Relationship Binding Model

The ability of a neural network to learn these DCC representations, and to systematically generalize to novel input patterns, was explored by O'Reilly and Busby (2002). This model demonstrates both that distributed, coarse-coded conjunctive representations can systematically perform binding relationships, and that not all mechanisms for developing such relationships are equivalent. The network (Figure 7a) was

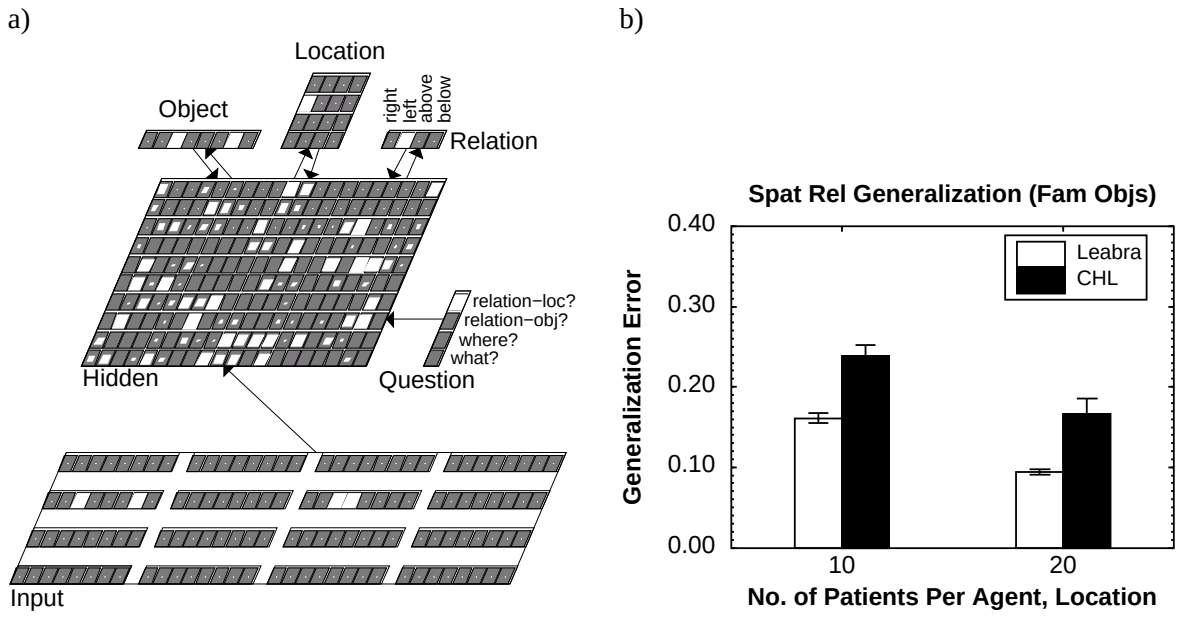


Figure 7: **a)** Spatial relationship binding model, representing posterior visual cortex (O'Reilly & Busby, 2002). Objects are represented by distributed patterns of activation over 8 feature values in each location, with the input containing a 4x4 array of object locations. Input patterns contain two different objects, arranged either vertically or horizontally. The network answers different questions about the inputs based on the activation of the Question input layer. For the "what?" question, the location of one of the objects is activated as an input in the Location layer, and the network must produce the correct object features for the object in that location. For the "where?" question, the object features for one of the objects are activated in the Object layer, and the network must produce the correct location activation for that object. For the "relation-obj?" question, the object features for one object are activated, and the network must activate the relationship between this object and the other object, in addition to activating the location for this object. For the "relation-loc?" question, the location of one of the objects is activated, and the network must activate the relationship between this object and the other object, in addition to activating the object features for this object (this is the example shown in the network, responding that the target object is to the left of the other object). Thus, the hidden layer must have bound object, location, and relationship information in its encoding of the input. **b)** Generalization results for different algorithms on the spatial relationship binding task (testing on familiar objects in novel locations; similar results hold for novel objects as well). Only the 400 Agent, Location x 10 or 20 Patient, Location cases are shown. It is clear that Leabra performed roughly twice as well as the CHL algorithm, consistent with earlier results on other tasks (O'Reilly, 2001).

trained to encode and report the spatial relationship between two items presented on its inputs, in addition to the identity and location of one of these items. Thus, the need for binding was taxed in two ways. First, the mere presence of two stimulus items demanded the ability to bind the features associated with one stimulus as distinct from the other. Second, and perhaps more challenging, the need to encode the spatial relationship information between objects required a kind of *relational* binding that has often been discussed in the context of complex structured knowledge representations (e.g., Touretzky, 1986; Hummel & Biederman, 1992; Hummel & Holyoak, 1997; Smolensky, 1990; Shastri & Ajjanagadde, 1993; Gasser & Colunga, 1998; Plate, 1995). Specifically, the network needed to be able to identify one of the two inputs as the "agent" item (i.e., the focus of attention), and report the relationship of the other "patient" item relative to it, and not the other way around.

The model is a very simplified rendition of the early visual system. During training the model is presented with a pair of input items in a simulated visual field, and is "asked" one of four corresponding questions (via the activation of a question input unit) (see Figure 7a for details). The model was implemented as a recurrent

neural network using the Leabra framework (O'Reilly & Munakata, 2000), and it achieved very high levels of generalization based on relatively limited amounts of experience (e.g., 95% correct after training on only 25% of the training space, and 80% correct after training on only roughly 10% of the space). In addition, a model using only contrastive Hebbian (CHL) error-driven learning, and another using the Almedia-Pineda recurrent backpropagation algorithm, were also run. Of these, it was found that Almedia-Pineda was not able to learn to successfully preform the task. While both the Leabra and CHL networks were able to learn, the additional constraints in Leabra (Hebbian learning and inhibitory competition) produced nearly twice as good generalization as CHL (Figure 7b).

Thus, by incorporating additional, biologically motivated constraints on the development of internal representations in the network, the Leabra model is able to achieve more systematicity in its representations, which subsequently give rise to better generalization performance. Importantly, we analyzed the internal representations of the Leabra network, and found that it developed both specialized representations of separable stimulus features (i.e., just representing what or where separately) and distributed coarse-coded conjunctive representations that integrated across features. This is typically what is observed in neural recording studies of the visual pathway, where many neurons encode strange conjunctions of stimulus features (Tanaka, 1996), while others have relatively more featural selectivity.

Other Mechanisms of Integration and Dissociation

There are numerous other neural mechanisms that can give rise over development to integration and dissociation of function within the cortex. These mechanisms are generally compatible with the above framework, but do not emerge directly from the overall computational tradeoffs behind it. A selection of such mechanisms are briefly reviewed here (see Jacobs, 1999 for a more detailed review).

It is well established that synapses proliferate early in development, and are then pruned as the brain matures (e.g., Huttenlocher, 1990). This process of refining the connectivity of neurons can lead to the development of more clearly delineated functional specializations in different brain areas (Johnson & Vecera, 1996), as has been demonstrated in computational models (Jacobs & Jordan, 1992; Miller, 1995). This process has been termed "parcellation". For example, Jacobs and Jordan (1992) showed that a network with a bias toward strengthening connections to physically proximal neurons produced a topographic organization of specialized functions within an initially homogeneous network. Although a focus on pruning is prevalent, others have emphasized the importance of the ongoing grown of new synapses, which can support continued plasticity of the system (Quartz & Sejnowski, 1997). As Jacobs (1999) points out, both pruning and synaptic growth behave functionally very similar to standard forms of Hebbian learning used in many different neural network models. Thus, it remains to be seen whether including these mechanisms in a broader range of models will result in fundamentally new computational properties. It could well be that these processes are a pragmatic physical necessity of wiring up the huge numbers of neurons in the mammalian cortex, whereas most small-scale models "cheat" and use full connectivity with Hebbian learning mechanisms, possibly with similar effect.

In both the parcellation models and Hebbian learning, competition plays a critical role in forcing the specialization of different neurons and brain areas. This competition can take place at many different scales, from synapses to neurons to larger-scale brain areas. This latter form of competition has been exploited in the *mixture of experts* models (Jacobs, Jordan, Nowlan, & Hinton, 1991; Jordan & Jacobs, 1994; Jacobs & Kosslyn, 1994). These models posit that learning is somehow gated by the relevance of a given group or pool of neurons (an "expert") for a given trial of learning. Experts that are most relevant get to learn the most from the feedback on a trial, and this causes further specialization of these experts for particular types of trials. Due to competition, experts for one set of trials typically lose out to other experts for other types of trials, resulting in an overall dissociation or specialization of function across these experts. This

may provide a reasonable computational model for specialization of function across different cortical areas. However, as noted in Jacobs (1999), it is unclear if the requisite large-scale competition between brain areas exists in the brain. Thus, it may make more sense to consider competition to operate fundamentally at the level of individual neurons (which is relatively well accepted), but to also allow for positive excitatory interactions among neurons. These excitatory interactions can cause neurons to group together and act as a more coherent whole. In effect, these excitatory grouping effects, together with pervasive inhibition mediated by local inhibitory interneurons, may result in emergent learning dynamics that resemble those captured in the mixture of experts models. This dynamic is present in several existing models of parcellation, for example in the development of ocular dominance columns (Miller, 1995).

In addition, these kinds of emergent competitive dynamics may have an overlay of more biologically-determined changes in plasticity over development. For example, one model explored the effects of "trophic waves" of plasticity that spread from simulated primary sensory areas to higher-level association areas (Shrager & Johnson, 1996). This trophic wave effect led to greater levels of neural specialization, in particular to the development of more complex higher-order representations in the higher-level association cortex.

These mechanisms are compelling and should be included more widely into neural network learning models. It will be interesting to explore in future work the possible interactions between these types of mechanisms and the general tradeoff principles articulated earlier.

General Discussion

The general conclusions from the computational tradeoffs described above are summarized in the tripartite cognitive architecture pictured back in Figure 1. This architecture is composed of posterior cortex (PC), hippocampus (HC), and frontal cortex/basal ganglia (FC), with each component specialized for a specific computational function. The posterior cortex is specialized for slowly developing rich, overlapping distributed representations that encode the general structure of the world, and for using these representations to support inferential reasoning through spreading activation dynamics, among other functions. The hippocampus uses sparse distributed representations to avoid interference while rapidly learning about arbitrary novel conjunctions (e.g., episodes), and recurrent connectivity in CA3 of the hippocampus supports pattern completion (recall) of previously encoded patterns. The frontal cortex/basal ganglia system uses relatively isolated representations and intrinsic bistability to robustly maintain information in an active state, and the basal ganglia provides adaptive gating to selectively update these representations according to task demands.

These distinctions between functional areas do not align with stimulus content dimensions. In contrast, each area encodes largely the same kinds of content (e.g., different stimulus dimensions and abstractions, language representations, etc), but does so in a different way, with different computational affordances. This allows the binding problem to be avoided, as each area can use distributed coarse-coded representations to efficiently and systematically cover the space of bindings that need to be distinguished.

This architecture lies between the extremes of modularity and equipotentiality — it has elements of both. However, it is not just any kind of both, but rather a very particular kind of both that focuses on some factors as critical for driving specialization, and not others. This approach can be summarized with the following "recipe" for "discovering" dissociated functional areas:

1. Identify essential functions.
2. Identify their requisite neural mechanisms (using computational models).
3. If they are incompatible, separate brain areas are required.

Of course, each of these steps requires considerable elaboration and judgment to be applied successfully, but this at least serves to highlight the core of the logic behind the present work.

This recipe can be applied within posterior cortex, for example to help understand the nature of the specialization in the fusiform face area (FFA) (Kanwisher, 2000; Tarr & Gauthier, 2000). From the hippocampal modeling work, we know that sparse activity levels lead to pattern separation, and thus the ability to distinctly represent a large number of similar input patterns. The apparent ability of the FFA to support identification of highly similar subordinate category members (e.g., faces) would certainly be greatly facilitated by this kind of sparse activity. Thus, it may be that this is what is unique about this brain area relative to other areas of posterior cortex. Note that because this area does not need to also support pattern completion from partial cues in the same way that the hippocampal system does, it therefore does not require the full set of neural specializations present in the hippocampus. In any case, this view of FFA specialization is appealing in its biological simplicity (it is easy to see how such a simple parametric variation could be genetically coded, for example), and is consistent with the notion that this area can also be co-opted for other forms of subordinate category representation (Tarr & Gauthier, 2000).

In conclusion, this paper has hopefully stimulated some interest in the notion that a cognitive architecture defined in terms of computational tradeoffs, with each area integrating information using distributed coarse-coded conjunctive representations to avoid binding problems, may provide some useful understanding of complex patterns of behavior from development to the mature system.

References

Adini, Y., Sagi, D., & Tsodyks, M. (2002). Context-enabled learning in the human visual system. *Nature*, *415*, 790–792.

Aggleton, J. P., & Brown, M. W. (1999). Episodic memory, amnesia, and the hippocampal-anterior thalamic axis. *Behavioral and Brain Sciences*, *22*, 425–490.

Alexander, G. E., DeLong, M. R., & Strick, P. L. (1986). Parallel organization of functionally segregated circuits linking basal ganglia and cortex. *Annual Review of Neuroscience*, *9*, 357–381.

Aslin, C., Blake, R., & Chun, M. M. (2002). Perceptual learning of temporal structure. *Vision Research*, *42*, 3019–3030.

Baddeley, A. D. (1986). *Working memory*. New York: Oxford University Press.

Barone, P., & Joseph, J. P. (1989). Prefrontal cortex and spatial sequencing in macaque monkey. *Experimental Brain Research*, *78*, 447–464.

Bloom, P., & Markson, L. (1998). Capacities underlying word learning. *Trends in Cognitive Science*, *2*, 67–73.

Braver, T. S., & Cohen, J. D. (2000). On the control of control: The role of dopamine in regulating prefrontal function and working memory. In S. Monsell, & J. Driver (Eds.), *Control of cognitive processes: Attention and performance XVIII* (pp. 713–737). Cambridge, MA: MIT Press.

Burgess, N., & O'Keefe, J. (1996). Neuronal computations underlying the firing of place cells and their role in navigation. *Hippocampus*, *6*, 749–762.

Casey, B. J., Durston, S., & Fossella, J. A. (2001). Evidence for a mechanistic model of cognitive control. *Clinical Neuroscience Research*, *1*, 267–282.

Cer, D. M., & O'Reilly, R. C. (in press). Neural mechanisms of binding in the hippocampus and neocortex: Insights from computational models. In H. D. Zimmer, A. Mecklinger, & U. Lindenberger (Eds.), *Binding in memory*. Oxford: Oxford University 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.

Csibra, G., Davis, G., & Johnson, M. H. (2000). Gamma oscillations and object processing in the infant brain. *Science, 290*, 1582.

Dominey, P. F., & Georgieff, N. (1997). Schizophrenics learn surface but not abstract structure in a serial reaction time task. *Neuroreport, 8*, 2877.

Eichenbaum, H., H., Otto, T., & Cohen, N. J. (1994). Two functional components of the hippocampal memory system. *Behavioral and Brain Sciences, 17*(3), 449–518.

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.

Engel, A. K., Konig, P., Kreiter, A. K., Schillen, T. B., & Singer, W. (1992). Temporal coding in the visual cortex: New vistas on integration in the nervous system. *Trends in Neurosciences, 15*(6), 218–226.

Frank, M. J., Rudy, J. W., & O'Reilly, R. C. (2003). Transitivity, flexibility, conjunctive representations and the hippocampus: II. a computational analysis. *Hippocampus, 13*, 341–354.

Freedman, D. J., Riesenhuber, M., Poggio, T., & Miller, E. K. (2002). Visual categorization and the primate prefrontal cortex: Neurophysiology and behavior. *Journal of Neurophysiology, 88*, 929–941.

Furmanski, C. S., & Engel, S. A. (2000). Perceptual learning in object recognition: Object specificity and size invariance. *Vision Research, 40*, 473.

Gasser, M., & Colunga, E. (1998). *Where do relations come from?* (Technical Report 221). Bloomington, IN: Indiana University Cognitive Science Program.

Gilbert, C. D., Sigman, M., & Crist, R. E. (2001). The neural basis of perceptual learning. *Neuron, 31*, 681–697.

Gillund, G., & Shiffrin, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review, 91*, 1–67.

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences, 15*(1), 20–25.

Gray, C. M., Engel, A. K., Konig, P., & Singer, W. (1992). Synchronization of oscillatory neuronal responses in cat striate cortex — temporal properties. *Visual Neuroscience, 8*, 337–347.

Graybiel, A. M., & Kimura, M. (1995). Adaptive neural networks in the basal ganglia. In J. C. Houk, J. L. Davis, & D. G. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 103–116). Cambridge, MA: MIT Press.

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.

Hayne, H., Boniface, J., & Barr, R. (2000). The development of declarative memory in human infants: Age-related changes in deferred imitation. *Behavioral Neuroscience, 114*, 77.

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed representations. In D. E. Rumelhart, J. L. McClelland, & PDP Research Group (Eds.), *Parallel distributed processing. Volume 1: Foundations* (Chap. 3, pp. 77–109). Cambridge, MA: MIT Press.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*, 528–551.

Holdstock, J. S., Mayes, A. R., Roberts, N., Cezayirli, E., Isaac, C. L., O'Reilly, R. C., & Norman, K. A. (2002). Under what conditions is recognition spared relative to recall after selective hippocampal damage in humans? *Hippocampus*, *12*, 341–351.

Howe, M. L., & Courage, M. L. (1993). On resolving the enigma of infantile amnesia. *Psychological Bulletin*, *113*, 305–326.

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, *99*(3), 480–517.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological Review*, *104*(3), 427–466.

Huttenlocher, P. R. (1990). Morphometric study of human cerebral cortex development. *Neuropsychologia*, *28*(6), 517–527.

Ito, M., Westheimer, G., & Gilbert, C. D. (1998). Attention and perceptual learning modulate contextual influences on visual perception. *Neuron*, *20*, 1191.

Jacobs, R. A. (1999). Computational studies of the development of functionally specialized neural modules. *Trends in Cognitive Sciences*, *3*, 31–38.

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

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., & Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural Computation*, *3*, 79–87.

Jacobs, R. A., & Kosslyn, S. M. (1994). Encoding shape and spatial relations: The role of receptive field size in coordinating complementary representations. *Cognitive science*, *18*, 361–386.

Johnson, M. H., & Vecera, S. P. (1996). Cortical differentiation and neurocognitive development: the parcellation conjecture. *Behavioral Processes*, *36*, 195–212.

Jordan, M. I., & Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural Computation*, *6*(2), 181–214.

Kanwisher, N. (2000). Domain specificity in face perception. *Nature Neuroscience*, *3*, 759–763.

Lambon-Ralph, M. A., Patterson, K., Garrard, P., & Hodges, J. R. (2003). Semantic dementia with category specificity: A comparative case-series study. *Cognitive Neuropsychology*, *20*, 307–326.

Lee, I., Yoganarasimha, D., Rao, G., & Knierim, J. J. (2004). Comparison of population coherence of place cells in hippocampal subfields CA1 and CA3. *Nature*, *430*, 456–459.

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.

Lewis, D. A. (1997). Development of the prefrontal cortex during adolescence: Insights into vulnerable neural circuits in schizophrenia. *Neuropsychopharmacology*, *16*, 385–398.

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.

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience, 4*, 310–322.

McNaughton, B. L., & Morris, R. G. M. (1987). Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends in Neurosciences, 10*(10), 408–415.

Mel, B. A., & Fiser, J. (2000). Minimizing binding errors using learned conjunctive features. *Neural Computation, 12*, 731–762.

Middleton, F. A., & Strick, P. L. (2000). Basal ganglia and cerebellar loops: Motor and cogntive circuits. *Brain Research Reviews, 31*, 236–250.

Miller, K. D. (1995). Receptive fields and maps in the visual cortex: Models of ocular dominance and orientation columns. In E. Domany, J. L. van Hemmen, & K. Schulten (Eds.), *Models of neural networks, III* (pp. 55–78). New York, NY: Springer Verlag.

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

Morton, J. B., & Munakata, Y. (2002a). Active versus latent representations: A neural network model of perseveration and dissociation in early childhood. *Developmental Psychobiology, 40*, 255–265.

Morton, J. B., & Munakata, Y. (2002b). Are you listening? Exploring a knowledge action dissociation in a speech interpretation task. *Developmental Science, 5*, 435–440.

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

Munakata, Y. (2004). Computational cognitive neuroscience of early memory development. *Developmental Review, 24*, 133–153.

Munakata, Y., & Yerys, B. E. (2001). All together now: When dissociations between knowledge and action disappear. *Psychological Science, 12*, 335–337.

Nieder, A., Freedman, D. J., & Miller, E. K. (2002). Representation of the quantity of visual items in the primate prefrontal cortex. *Science, 298*, 1708–1711.

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

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford, England: Oxford University Press.

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., & Busby, R. S. (2002). Generalizable relational binding from coarse-coded distributed representations. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems (NIPS) 14*. Cambridge, MA: MIT Press.

O'Reilly, R. C., Busby, R. S., & Soto, R. (2003). Three forms of binding and their neural substrates: Alternatives to temporal synchrony. In A. Cleeremans (Ed.), *The unity of consciousness: Binding, integration, and dissociation* (pp. 168–192). Oxford: Oxford University Press.

O'Reilly, R. C., & Frank, M. J. (submitted). Making working memory work: A computational model of learning in the frontal cortex and basal ganglia.

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., & Norman, K. A. (2002). Hippocampal and neocortical contributions to memory: Advances in the complementary learning systems framework. *Trends in Cognitive Sciences*, *6*, 505–510.

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

O'Reilly, R. C., & Soto, R. (2002). A model of the phonological loop: Generalization and binding. In T. G. Dietterich, S. Becker, & Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems (NIPS) 14*. Cambridge, MA: MIT Press.

Plate, T. A. (1995). Holographic reduced representations. *IEEE Transactions on Neural Networks*, *6*, 623–641.

Quartz, S. R., & Sejnowski, T. J. (1997). the neural basis of cognitive development: a constructivist manifesto. *the Behavioral and Brain Sciences*, *20*, 537.

Rao, S. C., Rainer, G., & Miller, E. K. (1997). Integration of what and where in the primate prefrontal cortex. *Science*, *276*, 821–824.

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.

Rolls, E. T. (1989). Functions of neuronal networks in the hippocampus and neocortex in memory. In J. H. Byrne, & W. O. Berry (Eds.), *Neural models of plasticity: Experimental and theoretical approaches* (pp. 240–265). San Diego, CA: Academic Press.

Rougier, N. P., Noelle, D., Braver, T. S., Cohen, J. D., & O'Reilly, R. C. (submitted). Prefrontal cortex and the flexibility of cognitive control: Rules without symbols.

Rudy, J. W., & O'Reilly, R. C. (2001). Conjunctive representations, the hippocampus, and contextual fear conditioning. *Cognitive, Affective, and Behavioral Neuroscience*, *1*, 66–82.

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.

Shadlen, M. N., & Movshon, J. A. (1999). Synchrony unbound: A critical evaluation of the temporal binding hypothesis. *Neuron*, *24*, 67–77.

Shastri, L., & Ajjanagadde, V. (1993). From simple associations to systematic reasoning: A connectionist representation of rules, variables, and dynamic bindings using temporal synchrony. *Behavioral and Brain Sciences*, *16*, 417–494.

Sherry, D. F., & Schacter, D. L. (1987). The evolution of multiple memory systems. *Psychological Review*, *94*(4), 439–454.

Shrager, J., & Johnson, M. H. (1996). Dynamic plasticity influences the emergence of function in a simple cortical array. *Neural Networks, 9*, 1119.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist networks. *Artificial Intelligence, 46*, 159–216.

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

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

Stark, C. E. L., & McClelland, J. L. (2000). Repetition priming of words, pseudowords, and nonwords. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26*, 945.

Sutherland, R. J., & Rudy, J. W. (1989). Configural association theory: The role of the hippocampal formation in learning, memory, and amnesia. *Psychobiology, 17*(2), 129–144.

Tanaka, K. (1996). Inferotemporal cortex and object vision. *Annual Review of Neuroscience, 19*, 109–139.

Tarr, M. J., & Gauthier, I. (2000). FFA: a flexible fusiform area for subordinate-level visual processing automatized by expertise. *Nature Neuroscience, 3*, 764–770.

Teyler, T. J., & Discenna, P. (1986). The hippocampal memory indexing theory. *Behavioral Neuroscience, 100*, 147–154.

Touretzky, D. S. (1986). BoltzCONS: Reconciling connectionism with the recursive nature of stacks and trees. *Proceedings of the 8th Annual Conference of the Cognitive Science Society* (pp. 522–530). Hillsdale, NJ: Lawrence Erlbaum Associates.

Touretzky, D. S., & Redish, A. D. (1996). A theory of rodent navigation based on interacting representations of space. *Hippocampus, 6*, 247–270.

Treves, A., & Rolls, E. T. (1994). A computational analysis of the role of the hippocampus in memory. *Hippocampus, 4*, 374–392.

Ungerleider, L. G., & Mishkin, M. (1982). Two cortical visual systems. In D. J. Ingle, M. A. Goodale, & R. J. W. Mansfield (Eds.), *The analysis of visual behavior*. Cambridge, MA: MIT Press.

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.

Vazdarjanova, A., & Guzowski, J. F. (in press). Differences in hippocampal neuronal population responses to modifications of an environmental context: Evidence for distinct, yet complementary, functions of CA3 and CA1 ensembles. *Journal of Neuroscience*.

von der Malsburg, C. (1981). The correlation theory of brain function. MPI Biophysical Chemistry, Internal Report 81-2. In E. Domany, J. L. van Hemmen, & K. Schulten (Eds.), *Models of neural networks, II (1994)*. Berlin: Springer.

Wagner, A. D., Koutstaal, W., Maril, A., Schacter, D. L., & Buckner, R. L. (2000). Task-specific repetition priming in left inferior prefrontal cortex. *Cerebral Cortex, 10*, 1176–1184.

Walker, G. A., Ohzawa, I., & Freeman, R. D. (1999). Asymmetric suppression outside the classical receptive field of the visual cortex. *Journal of Neuroscience, 19*, 10536.

Wallis, J. D., Anderson, K. C., & Miller, E. K. (2001). Single neurons in prefrontal cortex encode abstract rules. *Nature, 411*, 953–956.

Wickelgren, W. A. (1969). Context-sensitive coding, associative memory, and serial order in (speech) behavior. *Psychological Review*, *76*, 1–15.

Wu, X., Baxter, R. A., & Levy, W. B. (1996). Context codes and the effect of noisy learning on a simplified hippocampal CA3 model. *Biological Cybernetics*, *74*, 159–165.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*, 441–517.

Zelazo, P. D., Frye, D., & Rapus, T. (1996). An age-related dissociation between knowing rules and using them. *Cognitive Development*, *11*, 37–63.

Zemel, R. S., Williams, C. K., & Mozer, M. C. (1995). Lending direction to neural networks. *Neural Networks*, *8*, 503.