The Division of Labor Between the Neocortex and Hippocampus

Randall C. O'Reilly
Department of Psychology
University of Colorado Boulder
345 UCB
Boulder, CO 80309
oreilly@psych.colorado.edu

Abstract:

This chapter presents an overview of a computational approach towards understanding the different contributions of the neocortex and hippocampus in learning and memory. The approach is based on a set of principles derived from converging biological, psychological, and computational constraints. The most central principles are that the neocortex employs a slow learning rate and overlapping distributed representations to extract the general statistical structure of the environment, while the hippocampus learns rapidly using separated representations to encode the details of specific events while suffering minimal interference. Additional principles concern the nature of learning (error-driven and Hebbian), and recall of information via pattern completion. These principles are demonstrated through neocortical and hippocampal models of a well-known memory task, the AB-AC paired associates task. The results of applying these principles to a wide range of phenomena in conditioning, habituation, contextual learning, recognition memory, recall, and retrograde amnesia, are also summarized.

Introduction

In addition to addressing specific patterns of behavioral and neural data, neural network models are valuable for their ability to establish general principles of functional neural organization. In particular, computational models can explain how differences in the structuring and parameters of neural networks can lead to qualitatively different, often mutually incompatible capabilities. To the extent that different computational capacities require fundamentally different kinds of neural networks, the brain could either have a *compromise* or *tradeoff* between the different network properties, or it could *specialize* different brain areas for different functions to avoid such a tradeoff. Critically, this kind of computational approach to functional neural organization enables one to understand both *what* is different about the way different neural systems learn, and *why*, from a functional perspective, they should have these differences in the first place. Thus, the computational approach can go beyond mere description towards understanding the deeper principles underlying the organization of the cognitive system.

This chapter explores computational tradeoffs to understand the functional organization of brain areas involved in learning and memory, specifically the hippocampus and various areas of the neocortex. These tradeoffs are based on a set of computational principles, derived from a convergence of biological, psychological, and computational constraints, for understanding how neural systems subserve learning and memory. These principles were first developed in McClelland, McNaughton, and O'Reilly (1995), and have been refined several times since then (O'Reilly, Norman, & McClelland, 1998; O'Reilly & Rudy, 2000, 2001; O'Reilly & Munakata, 2000). The computational principles have been applied to a wide range of learning and memory phenomena across several species (rats, monkeys and humans). For example, they can account for impaired and preserved learning capacities with hippocampal lesions in conditioning, habituation, contextual learning, recognition memory, recall, and retrograde amnesia.

The chapter begins with an concise exposition of the principles, adapted from O'Reilly and Rudy (2000), and then discusses in somewhat more detail a set of specific simulations from O'Reilly and Munakata (2000) that illustrate these principles. It concludes with an overview of recent and ongoing applications of the models to a variety of cognitive neuroscience phenomena.

The Principles

Several levels of principles are presented, building from those that are most obvious from a mechanistic perspective. Only basic neural network mechanisms are needed to understand these principles. Specifically, we

Chapter for Book on Connectionist Modeling in Cognitive (Neuro-)Science, George Houghton, Ed., Psychology Press. Supported by ONR grants N00014-00-1-0246 and N00014-03-1-0428, and NIH grants MH061316-01 and MH64445.

assume networks having a number of units that communicate via propagation of activation signals along weighted connections to other units. Furthermore, we assume that learning occurs through changes to the connection weights, through widely-known mechanisms as discussed below. See chapter ?? of this volume for a review of these basic mechanisms, and O'Reilly and Munakata (2000) for a recent, in-depth treatment that covers the biological bases of these mechanisms.

Learning Rate, Overlap, and Interference

The most basic set of principles can be motivated by considering how subsequent learning can interfere with prior learning. A classic example of this kind of interference can be found in the $AB - AC$ associative learning task (e.g., Barnes & Underwood, 1959). The A represents one set of words that are associated with two different sets of other words, B and C. For example, the word *window* will be associated with the word *reason* in the AB list, and associated with *locomotive* on the AC list. After studying the AB list of associates, subjects are tested by asking them to give the appropriate B associate for each of the A words. Then, subjects study the AC list (often over multiple iterations), and are subsequently tested on both lists for recall of the associates after each iteration of learning the AC list. Subjects exhibit some level of interference on the initially learned AB associations as a result of learning the AC list, but they still remember a reasonable percentage (see Figure 1a for representative data).

The first set of principles concern the ability of a network to rapidly learn new information with a level of interference characteristic of human subjects, as in the $AB - AC$ task. Specifically, we consider the effects of *overlapping representations* and *rate of learning*. Overlapping representations arise in distributed patterns of neural activity over multiple units, where subsets of units are shared across different representations. For example, one would imagine that there are shared units across the AB and AC patterns in the $AB - AC$ task (deriving from the shared A element). Rate of learning refers to the size of weight changes made during learning. Both of these factors affect interference as follows:

- Overlapping representations lead to interference (conversely, separated representations prevent interference).
- A faster learning rate causes more interference (conversely, a slower learning rate causes less interference).

The mechanistic basis for these principles within a neural network perspective is straightforward. Interference is

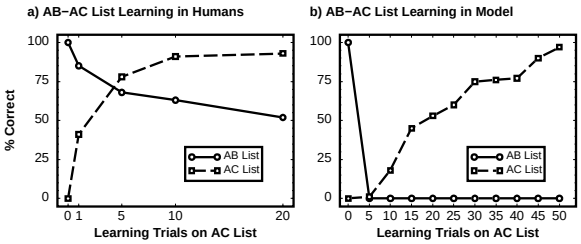

Figure 1: Human and model data for AB-AC list learning. a) Humans show some interference for the AB list items as a function of new learning on the AC list items. b) Model shows a catastrophic level of interference. (data reproduced from McCloskey & Cohen, 1989).

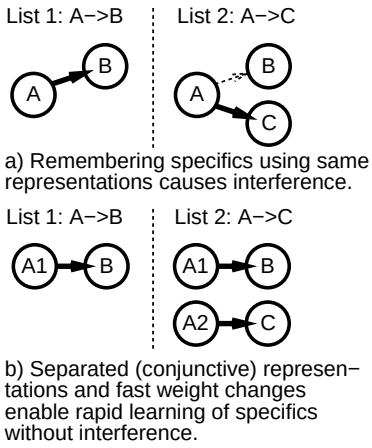

Figure 2: Interference as a function of overlapping (same) representations versus separated representations. **a)** Using the same representation to encode two different associations ($A \rightarrow B$ and $A \rightarrow C$) causes interference — the subsequent learning of $A \rightarrow C$ interferes with the prior learning of $A \rightarrow B$ because the A stimulus must have stronger weights to C than to B for the second association, as is reflected in the weights. **b)** A separated representation, where A is encoded separated for the first list ($A1$) versus the second list ($A2$) prevents interference.

caused when weights used to encode one association are disturbed by the encoding of another (Figure 2a). Overlapping patterns share more weights, and therefore lead to greater amounts of interference. Clearly, if entirely separate representations are used to encode two different associations, then there will be no interference whatsoever (Figure 2b). The story with learning rate is similarly straightforward. Faster learning rates lead to more weight change, and thus greater interference (Figure 3). However, a fast learning rate is necessary for rapid learning.

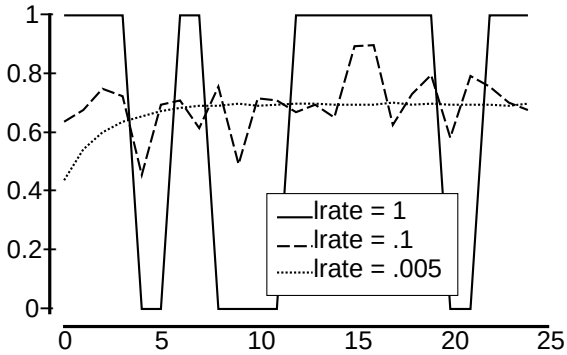

Figure 3: Weight value learning about a single input unit that is either active or not. The weight increases when the input is on, and decrease when it is off, in proportion to the size of the learning rate. The input has an overall probability of being active of .7. Larger learning rates (.1 or 1) lead to more interference on prior learning, resulting in a weight value that bounces around substantially with each training example. In the extreme case of a learning rate of 1, the weight only reflects what happened on the previous trial, retaining no memory for prior events at all. As the learning rate gets smaller (.005), the weight smoothly averages over individual events and reflects the overall statistical probability of the input being active.

Integration and Extracting Statistical Structure

Figure 3 shows the flip side of the interference story, *integration*. If the learning rate is low, then the weights will integrate over many experiences, reflecting the *underlying statistics* of the environment (White, 1989; McClelland et al., 1995). One can think of this in terms of computing a statistical average — each term in the average contributes by a small factor of $\frac{1}{N}$. As N gets larger, each term contributes less. Similarly, if a neural network is to compute the average or expected characteristics of the environment, each experience must contribute something like $\frac{1}{N}$ (where N is the total number of experiences) to the overall representation. Furthermore, overlapping representations facilitate this integration process, because the same weights need to be reused across many different experiences to enable the integration produced by a slow learning rate. This leads to the next principle:

- Integration across experiences to extract underlying statistical structure requires a slow learning rate and overlapping representations.

Episodic Memory and Generalization: Incompatible Functions

Thus, focusing only on pattern overlap for the moment, we can see that networks can be optimized for two different, and incompatible, functions: avoiding interference or integrating across experiences to extract generalities. Avoiding interference requires separated representations, while integration requires overlapping representations. These two functions each have clear functional advantages, leading to a further set of principles:

- Interference avoidance is essential for *episodic* memory, which requires learning about the specifics of individual events and keeping them separate from other events.
- Integration is essential for encoding the general statistical structure of the environment, abstracted away from the specifics of individual events, which enables *generalization* to novel situations.

The incompatibility between these functions is further evident in these descriptions (i.e., encoding specifics versus abstracting away from them). Also, episodic memory requires relatively rapid learning — an event must be encoded as it happens, and does not typically repeat itself for further learning opportunities. This completes a pattern of opposition between these functions: episodic learning requires rapid learning while integration and generalization requires slow learning. This is summarized in the following principle:

- Episodic memory and extracting generalities are in opposition. Episodic memory requires rapid learning and separated patterns, while extracting generalities requires slow learning and overlapping patterns.

Applying the Principles to the Hippocampus and Neocortex

Armed with these principles, the finding that neural network models that have highly overlapping representations exhibit *catastrophic* levels of interference (McCloskey & Cohen, 1989, Figure 1b) should not be surprising. A number of researchers showed that this interference can be reduced by introducing various factors that result in less pattern overlap (e.g., Kortge, 1993; French, 1992; Sloman & Rumelhart, 1992; McRae & Hetherington, 1993). Thus, instead of concluding that neural networks may be fundamentally flawed, as McCloskey and Cohen (1989) argued (and a number of others have uncritically accepted), McClelland et al. (1995) argued that this catastrophic failure serves as an important clue into the structure of the human brain.

Specifically, we argued that because of the fundamental incompatibility between episodic memory and extracting generalities, the brain should employ two separate systems that each optimize these two objectives individually, instead of having a single system that tries to strike an inferior compromise:

- The hippocampus rapidly binds together information using pattern-separated representations to minimize interference.
- The neocortex slowly learns about the general statistical structure of the environment using overlapping distributed representations.

(see also Sherry & Schacter, 1987 for a similar conclusion). This line of reasoning provides a strikingly good fit to the known properties of the hippocampus and neocortex, respectively (see O'Reilly & Rudy, 2001; Norman & O'Reilly, in press; O'Reilly et al., 1998 for some examples).

Sparse Conjunctive Representations

The *conjunctive* or *configural* representations theory provides a converging line of thinking about the nature of hippocampal function (Sutherland & Rudy, 1989; Rudy & Sutherland, 1995; Wickelgren, 1979; O'Reilly & Rudy, 2001). A conjunctive/configural representation is one that binds together (conjoins or configures) multiple elements into a novel unitary representation. This is consistent with the description of hippocampal function given above, based on the need to separate patterns to avoid interference. Indeed, it is clear that pattern separation and conjunctive representations are two sides of the same coin, and that both are caused by the use of *sparse* representations (having relatively few active neurons) that are a known property of the hippocampus (O'Reilly & McClelland, 1994; O'Reilly & Rudy, 2001).

To understand why a sparse representation can lead to pattern separation (using different neurons to encode different representations), and conjunctivity, we consider two related explanations. First, consider a situation where the hippocampal representation is generated at random with some fixed probability of a unit becoming active. In this case, if fewer units are active, the odds that the same units will be active in two different patterns will go down (Figure 4). For example, if the probability of becoming active for one pattern (i.e., the sparseness) is .25, then the probability of becoming active for both patterns would be $.25^2$ or .0625. If the patterns are made more sparse so that the probability is now .05 for being active in one pattern, the probability of being active in both patterns falls to .0025. Thus, the pattern overlap is reduced by a factor of 25 by reducing the sparseness by a factor of 5 in this case. However, this analysis does not capture the entire story because it fails to take into account the fact that hippocampal units are actually driven by weighted connections with the input patterns, and therefore will be affected by similarity (overlap) in the input.

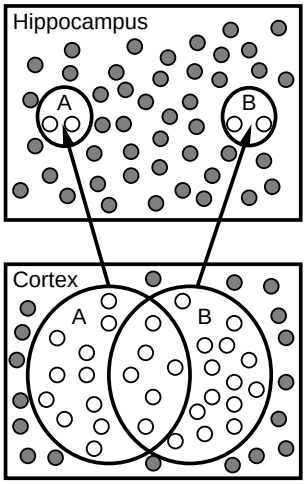

Figure 4: Pattern separation in the hippocampus. Small circles represent units, with active ones in white, inactive ones in grey. Circles A and B in the cortex and hippocampus indicate two sets of representations composed of patterns of active units. In the cortex, they are overlapping, and encompass relatively large proportion of active units. In the hippocampus, the representations are sparser as indicated by their smaller size, and thus overlap less (more pattern separation). Also, units in the hippocampus are conjunctive and are activated only by specific combinations of activity in the cortex.

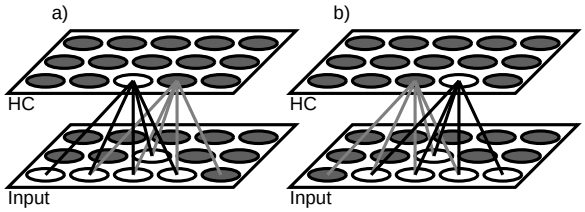

Figure 5: Conjunctive, pattern-separated representations result from sparseness (active units are represented in white, inactive ones in grey). The extreme case where only one receiving unit (in the upper layer, representing the hippocampus) is allowed to be active is shown here for simplicity. Each receiving unit has roughly the same number of randomly distributed connections from the input units. The two shown here have overlapping input connections, except for one unique unit each. Thus, two very similar input patterns sharing all the overlapping units and differing only in these unique units (shown in panels a and b) will get completely non-overlapping (separated) memory representations. In this way, the conjunctive memory representation resulting from sparseness produces pattern separation.

A more complete understanding of pattern separation can be achieved by considering the concept of a unit's *activation threshold* — how much excitation it requires to overcome the inhibitory competition from other units (Marr, 1969; O'Reilly & McClelland, 1994). To produce sparse representations, this threshold must be relatively high (e.g., because the level of inhibition is relatively strong for a given amount of excitatory input). Figure 5 shows how a high inhibitory threshold leads simultaneously to both pattern separation and conjunctive representations, where the hippocampal units depend critically on the conjunction of active units in the input. A high threshold leads to conjunctive representations because only those units having the closest alignment of their weight patterns with the current input activity pattern will receive enough excitation to become activated. In other words, the activation a unit receives must be a relatively high proportion of the total number of input units that are active, meaning that it is the specific combination or conjunction of these inputs that are responsible for driving the units. Figure 5 illustrates this effect in the extreme case where only the most excited receiving unit gets active. In reality, multiple (roughly 1-5%) units are activated in the hippocampus at any given time, but the same principle applies (see O'Reilly & McClelland, 1994 for a detailed analysis).

To summarize:

- Sparse hippocampal representations lead to pattern separation (to avoid interference) and conjunctive representations (to bind together features into a unitary representation).

This principle will be explored in greater detail in the context of a specific simulation described below.

Pattern Completion: Recalling a Conjunction

Pattern completion is required for recalling information from conjunctive hippocampal representations, yet it conflicts with the process of pattern separation that forms these representations in the first place (O'Reilly & McClelland, 1994). Pattern completion occurs when a partial input cue drives the hippocampus to complete to an entire previously-encoded set of features that were bound together in a conjunctive representation. For a given input pattern, a decision must be made to recognize it as a retrieval cue for a previous memory and perform pattern completion, or to perform pattern separation and store the input as a new memory. This decision is often difficult given noisy inputs and degraded memories. The hippocampus implements this decision as the effects of a set of basic mechanisms operating on input patterns (O'Reilly & McClelland, 1994; Hasselmo & Wyble, 1997), and it does not always do what would seem to be the right thing to do from an omniscient perspective knowing all the relevant task factors.

Learning Mechanisms: Hebbian and Error Driven

To more fully explain the roles of the hippocampus and neocortex we need to understand how learning works in these systems (the basic principles just described do not depend on the detailed nature of the learning mechanisms; White, 1989). There are two basic mechanisms that have been discussed in the literature, Hebbian and error-driven learning (e.g., Marr, 1971; McNaughton & Morris, 1987; Gluck & Myers, 1993; Schmajuk & DiCarlo, 1992). Briefly, Hebbian learning (Hebb, 1949) works by increasing weights between co-active neurons (and usually decreasing weights when a receiver is active and the sender is not), which is a well-established property of biological synaptic modification mechanisms (e.g., Collingridge & Bliss, 1987). Hebbian learning is useful for binding together features active at the same time (e.g., within the same episode), and has therefore been widely suggested as a hippocampal learning mechanism (e.g., Marr, 1971; McNaughton & Morris, 1987).

Error-driven learning works by adjusting weights to minimize the errors in a network's performance, with the best example of this being the *error backpropagation* algorithm (Rumelhart, Hinton, & Williams, 1986). Error-driven learning is sensitive to task demands in a way that Hebbian learning is not, and this makes it a much more capable form of learning for actually achieving a desired input/output mapping. Thus, it is natural to associate this form of learning with the kind of procedural or task-driven learning that the neocortex is often thought to specialize in (e.g., because amnesics with hippocampal damage have preserved procedural learning abilities). Although the backpropagation mechanism has been widely challenged as biologically implausible (e.g., Crick, 1989; Zipser & Andersen, 1988), a recent analysis shows that simple biologically-based mechanisms can be used to implement this mechanism (O'Reilly, 1996), so that it is quite reasonable to assume that the cortex depends on this kind of learning.

Although the association of Hebbian learning with the hippocampus and error-driven learning with the cortex is appealing in some ways, it turns out that both kinds of learning play important roles in both systems (O'Reilly & Rudy, 2001; O'Reilly & Munakata, 2000; O'Reilly, 1998). Thus, the specific learning principles adopted here are that both forms of learning operate in both systems:

- Hebbian learning binds together co-occurring features (in the hippocampus) and generally learns

about the co-occurrence statistics in the environment across many different patterns (in neocortex).

- Error-driven learning shapes learning according to specific task demands (shifting the balance of pattern separation and completion in the hippocampus, and developing task-appropriate representations in the neocortex).

It is the existence of this task-driven learning that complicates the picture for nonlinear discrimination learning problems.

A Summary of Principles

The above principles can be summarized with the following three general statements of neocortical and hippocampal learning properties (O'Reilly & Rudy, 2001; O'Reilly & Norman, 2002):

Learning rate. The cortical system typically learns slowly, while the hippocampal system typically learns rapidly.

Conjunctive bias. The cortical system has a bias towards integrating over specific instances to extract generalities. The hippocampal system is biased by its intrinsic sparseness to develop conjunctive representations of specific instances of environmental inputs. However, this conjunctive bias trades-off with the countervailing process of pattern completion, so the hippocampus does not always develop new conjunctive representations (sometimes it completes to existing ones).

Learning mechanisms. Both cortex and hippocampus use error-driven and Hebbian learning. The error-driven aspect responds to task demands, and will cause the network to learn to represent whatever is needed to achieve goals or ends. Thus, the cortex can overcome its bias and develop specific, conjunctive representations if the task demands require this. Also, error-driven learning can shift the hippocampus from performing pattern separation to performing pattern completion, or vice-versa, as dictated by the task. Hebbian learning is constantly operating, and reinforcing the representations that are activated in the two systems.

These principles are focused on distinguishing neocortex and hippocampus — we have also articulated a more complete set of principles that are largely common to both systems (O'Reilly, 1998; O'Reilly & Munakata, 2000). Models incorporating these principles have been extensively applied to a wide range of different cortical phenomena, including perception, language, and higher-level cognition. In the next sections, we explore these principles in the context of two implemented models, one representing a basic cortical network, and the other a hippocampal network based on the neurobiology of the hippocampal system. Both networks are tested on the $AB - AC$ interference task as described earlier.

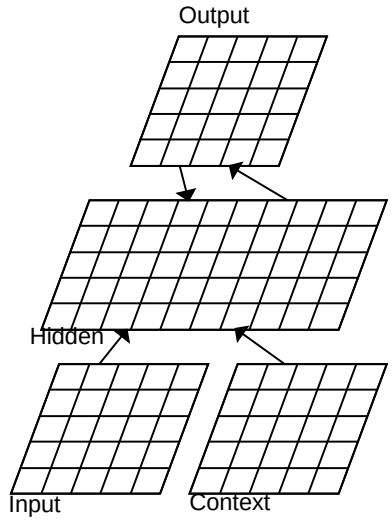

Figure 6: Network for the AB–AC list learning task, with the Input layer representing the A stimulus, the Context input representing the list context, and the Output being the B or C word associate, depending on the list context.

Neocortical Model of the $AB - AC$ Task

As we mentioned previously, the McCloskey and Cohen (1989) catastrophic interference model can be interpreted as a good example of what would happen to a neocortical network with distributed, overlapping representations on the $AB - AC$ task. In this section, we examine a similar such model, developed in O'Reilly and Munakata (2000), that enables us to explore the important parameter of the sparseness, and thereby the level of pattern separation or overlap in the network's representations. Therefore, this network provides a concrete demonstration of some of the central principles outlined above.

Basic Properties of the Model

The basic framework for implementing the AB–AC task is to have two input patterns, one that represents the A stimulus, and the other that represents the "list context" (figure 6). Thus, we assume that the subject develops some internal representation that identifies the two different lists, and that this serves as a means of dis-

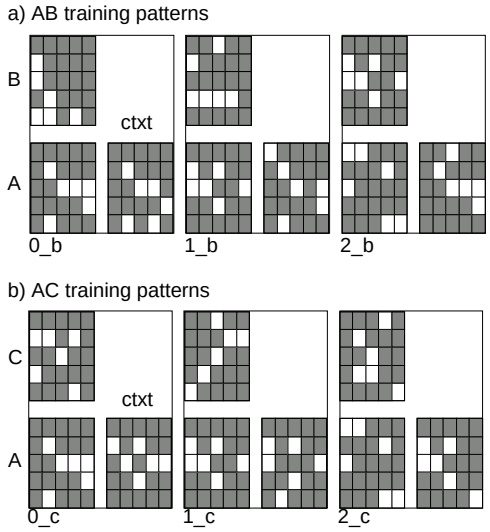

Figure 7: Training patterns for the AB–AC list learning task, showing the first 3 out of 10 patterns for the a) AB list and b) AC list. Notice that the A input is the same across both lists, but it is paired with a B output on the AB list, and a C output on the AC list. The context input is a random permutation of a different random pattern for each list.

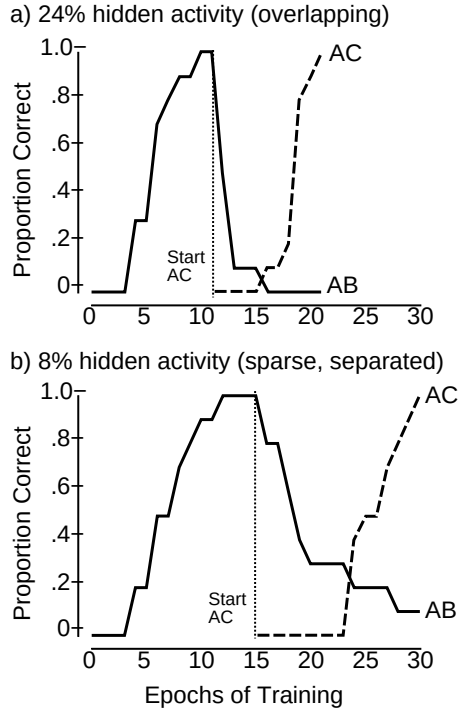

Figure 8: Example training results for two different parameter regimes. For both cases, the AB training is automatically terminated at 100% accuracy, switching to AC training. a) with 24% hidden activity and default parameters, the hidden representations are highly overlapping, and this produces rapid and complete interference on the AB list as the network starts learning AC. b) with sparser activity (8%) and other optimized parameters as described in the text, the onset of AB interference is slowed and asymptotic levels are reduced. However, these effects are not particularly dramatic, and do not match human performance.

ambiguating which of the two associates should be produced. These input patterns feed into a hidden layer, which then produces an output pattern corresponding to the B or C associate. We use a distributed representation of random bit patterns to represent the word stimuli (figure 7). The list context patterns for all the items on the AB list are all similar random variations of a common underlying pattern, and likewise for the AC items. Thus, the list context patterns are not identical for each item on the same list, just very similar (this reflects the fact that even if the external environmental context is constant, the internal perception of it fluctuates, and other internal context factors like a sense of time passing change as well).

To train the network, O'Reilly and Munakata (2000) initially used default parameters for the Leabra algorithm, which incorporates a comprehensive set of standard neural network mechanisms (O'Reilly & Munakata, 2000; O'Reilly, 1998, 2001). These mechanisms include a combination of both Hebbian and error-driven learning rules, and, most relevant for the present purposes, a k-winners-take-all (kWTA) inhibition function that can be used to explore different levels of representational sparseness. The default parameters specify a k value that produces an overall activation level of roughly 25%, meaning in the present network that 12 out of the 50 hidden units are active for any given input pattern (i.e., $k = 12$ winners). This activation level typically produces very good results on learning tasks typical of what we think the neocortex excels at — encoding efficient features for representing visual scenes, learning to recognize objects, learning to pronounce written words, etc. (see O'Reilly & Munakata, 2000 for many such models). All of these tasks involve extracting regularities over many instances of experience.

Figure 8a shows typical results with these standard parameters. You can see that just as the network starts to learn the AC list (after first acquiring the AB list), performance on the AB list deteriorates dramatically. This replicates the McCloskey and Cohen (1989) finding that a "generic" neural network does not do a good job of capturing the human data (figure 1). With the generic network as a baseline, we can now test the idea that different parameters can reduce the level of interference. The intention here is to illuminate the principles underlying these interference effects, and show how they could potentially be reduced — though we will see that they have relatively small effects in this particular context.

The primary source of interference in this network is the pattern overlap, resulting from the 25% activity level. Specifically, items on the AC list will activate and reuse the same units from the AB list, altering their weights to support the C associate instead of the B. Thus, by reducing the extent to which the hidden unit representations overlap (i.e., by making them *sparser*), we might be able to encourage the network to use separate representations for learning these two lists of items. To make the representations sparser, the k parameter in the kWTA function can be reduced to 4 instead of the default value of 12, resulting in an 8% overall activation level. Note that because the network depends on distributed representations for learning, one cannot reduce the activation much further without impairing learning performance.

O'Reilly and Munakata (2000) found that just reducing the k parameter to 4 did not actually make that much of a difference, because nothing was done to encourage it to use *different* sets of 4 units to represent the different associates. One way we can encourage this is to increase the variance of the initial random weights, making each unit have a more quirky pattern of responses that should encourage different units to encode the different associates. Another thing that can be done to improve performance is to enhance the contribution of the list context inputs relative to the A stimulus, because this list context disambiguates the two different associates. This can be achieved by a weight scaling parameter in Leabra, and it can be thought of as reflecting strategic focusing of attention by the subject. Finally, increased amounts of Hebbian learning might contribute to better performance because of the strong correlation of all items on a given list with the associated list context representation, which should be emphasized by Hebbian learning. This could lead to different subsets of hidden units representing the items on the two lists because of the different context representations. The balance of Hebbian and error-driven learning is controlled by a normalized scaling parameter, and the default level is .01 Hebbian and .99 error-driven — we can increase this to .05 Hebbian and .95 error-driven (note that the small weighting factor on Hebbian learning is typically used because Hebbian learning provides much larger and more consistent weight changes relative to error-driven).

Figure 8b shows typical results from the network with all the parameter changes just described — O'Reilly and Munakata (2000) found that these parameters produced the best results. The main results of these parameters were to delay the onset of interference, and to improve the final level of performance on the AB list after learning about AC. However, it is obvious that this network is still not performing at the level of human subjects, who still remember roughly 60% of the AB list after learning the AC list. In the next section, we show that the remaining limitations of the present network are probably due to the architecture of the network, because a network with an architecture based on the biology of the hippocampus is able to perform at the level of human subjects on this task.

	Rat		Model	
Area	Neurons	Activity (pct)	Units	Activity (pct)
EC	200,000	7.0	96	25.0
DG	1,000,000	0.5	250	1.6
CA3	160,000	2.5	160	6.25
CA1	250,000	2.5	256	9.4

Table 1: Rough estimates of the size of various hippocampal areas and their expected activity levels in the rat, and corresponding values in the model. Rat data from Squire et al., 1989; Boss et al., 1987; Boss et al., 1985; Barnes et al., 1990.

Hippocampal Model of the $AB - AC$ Task

The hippocampal formation has a distinctive and relatively well-known anatomical structure. Furthermore, considerable neural recording data has been obtained from the hippocampus, providing information about important parameters such as the relative activation levels in different areas of the hippocampus (table 1). The somewhat remarkable thing about the model described here is that by incorporating these features of the hippocampal biology, we find that the resulting model performs quite well on rapid learning tasks like the $AB - AC$ task, without suffering catastrophic levels of interference (O'Reilly & Munakata, 2000). A complete explication of the computational features of the biological structure of the hippocampus is beyond the scope of this chapter (see O'Reilly & McClelland, 1994 for one such treatment), but a few of the main points are covered here, followed by a description of how the model performs on the $AB - AC$ task.

Basic Properties of the Model

The model is based on what McNaughton has termed the "Hebb-Marr" model of hippocampal function (Hebb, 1949; Marr, 1969, 1970, 1971; McNaughton & Morris, 1987; McNaughton & Nadel, 1990). This model provides a framework for associating functional properties of memory with the biological properties of the hippocampus. Under this model, the two basic computational structures in the hippocampus are the feedforward pathway from the entorhinal cortex (EC) to area CA3 (via the dentate gyrus, DG), which is important for establishing the encoding of new memories, and the recurrent connectivity within CA3, which is primarily important for recalling previously stored memories.

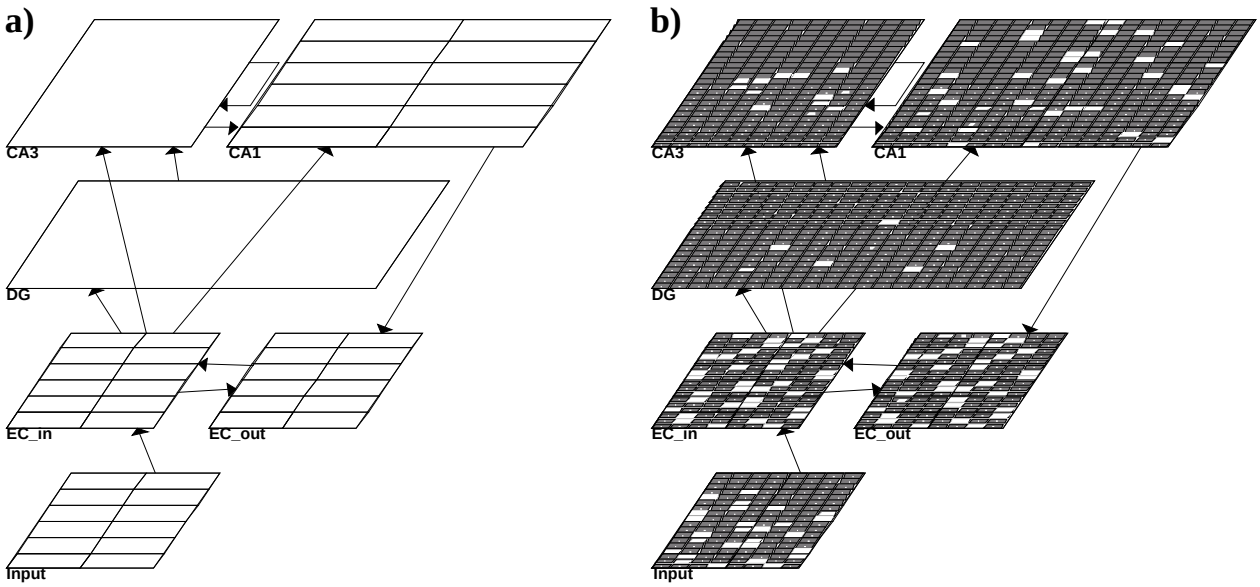

Figure 9: The hippocampus model. **a)** The areas and connectivity, and the corresponding columns within the Input, EC, and CA1. **b)** An example activity pattern. Note the sparse activity in the DG and CA3, and intermediate sparseness of the CA1.

Figure 9 shows the structure of the model, and an example activation pattern (see O'Reilly & Munakata, 2000 for the original presentation of this specific model, and O'Reilly et al., 1998; O'Reilly & Rudy, 2001 for similar ones). Table 1 shows that the model layers are roughly proportionately scaled based on the anatomy of the rat, but the activation levels are generally higher (less sparse) to obtain sufficient absolute numbers of active units for reasonable distributed representations given the small total number of units. The model is implemented using the same basic Leabra mechanisms as the earlier cortical model, with the activity levels enforced by setting appropriate k parameters in the kWTA inhibition function. Only Hebbian learning is used because it is sufficient for simple information storage, but it is likely that the hippocampus can also take advantage of error-driven learning in more complex tasks (O'Reilly & Rudy, 2001).

We can summarize the basic operations of the model by explaining how the encoding and retrieval of memories works in terms of the areas and projections of the hippocampus. The general scheme for encoding is that activation comes into the EC from the cortex, and then flows to the DG and CA3, forming a pattern separated representation across a sparse, distributed set of units that are then bound together by rapid Hebbian learning within the recurrent collaterals (also, learning in the feedforward pathway helps to encode the representation). Simultaneously, activation flows from the EC to the CA1, forming a somewhat pattern separated but also *invertible* representation there — that is, the CA1 representation can be inverted to reinstate the corresponding pattern of activity over the EC that originally gave rise to the CA1 pattern in the first place (McClelland & Goddard, 1996). An association between the CA3 and CA1 representations is encoded by learning in the connections between them.

Having encoded the information in this way, retrieval from a partial input cue can occur as follows. Again, the EC representation of the partial cue (based on inputs from the cortex) goes up to the DG and CA3. Now, the prior learning in the feedforward pathway and the recurrent CA3 connections leads to the ability to complete this partial input cue and recover the original CA3 representation. This completed CA3 representation then activates the corresponding CA1 representation, which, because it is invertible, is capable of recreating the complete original EC representation.

If, on the other hand, the EC input pattern is novel, then the weights will not have been facilitated for this particular activity pattern, and the CA1 will not be strongly driven by the CA3. Even if the EC activity pattern corresponds to two components that were previously studied, but not together, the conjunctive nature of the CA3 representations will prevent recall (O'Reilly et al., 1998).

In addition to capturing the rough sizes of the different hippocampal areas, The model incorporates rough approximations of the detailed patterns of connectivity within the hippocampal areas (e.g., Squire et al., 1989). The *perforant path* projections from EC to DG and CA3 are broad and diffuse, but the projection between the DG and CA3, known as the mossy fiber pathway, is sparse,

focused, and topographic. Each CA3 neuron receives only around 52-87 synapses from the mossy fiber projection in the rat, but it is widely believed that each synapse is significantly stronger than the perforant path inputs to CA3. In the model, each CA3 unit receives from 25% of the EC, and 10% of the DG. The lateral (recurrent) projections within the CA3 project widely throughout the CA3, and a given CA3 neuron will receive from a large number of inputs sampled from the entire CA3 population. Similarly, the Schaffer collaterals, which go from the CA3 to the CA1, are diffuse and widespread, connecting a wide range of CA3 to CA1. In the model, these pathways have full connectivity. Finally, the interconnectivity between the EC and CA1 is relatively point-to-point, not diffuse like the projections from EC to DG and CA3 (Tamamaki, 1991). This is captured in the model by a columnar structure and connectivity of CA1.

The functional properties of these various connectivity patterns have been analyzed (see O'Reilly & McClelland, 1994; McClelland & Goddard, 1996; O'Reilly et al., 1998 for details). We just focus on one example here, which is the broad and diffuse nature of the perforant pathway connectivity between EC and DG, CA3. This connectivity produces the same kind of effect as increasing the variance of weights that was explored in the neocortical network — it ensures that individual DG and CA3 neurons receive from a broadly distributed, and essentially random, subset of inputs. Thus, these neurons will encode different random conjunctions, and because of the extreme competition due to sparse activation levels, different such conjunctive units will be activated for even relatively similar input patterns. This produces pattern separation, and thereby avoids the interference problems that plague the neocortical network.

Performance on the $AB - AC$ Task

Now we describe how all of this hippocampal circuitry, as captured in the O'Reilly and Munakata (2000) model, performs on the AB–AC paired associates list learning task. If this circuitry enables the hippocampus to learn rapidly using pattern-separated representations that avoid interference, the model should be able to learn the new paired associates (AC) without causing undue levels of interference to the original AB associations, and it should be able to do this much more rapidly than was possible in the cortical model.

The model is trained in much the same way as the cortical model was. During training, the input patterns presented to the *Input* layer of the network were composed of 3 components. The first and second components represented the A and B or C associates, respectively, while the third was a representation of the list context. The item representations were simple random bit

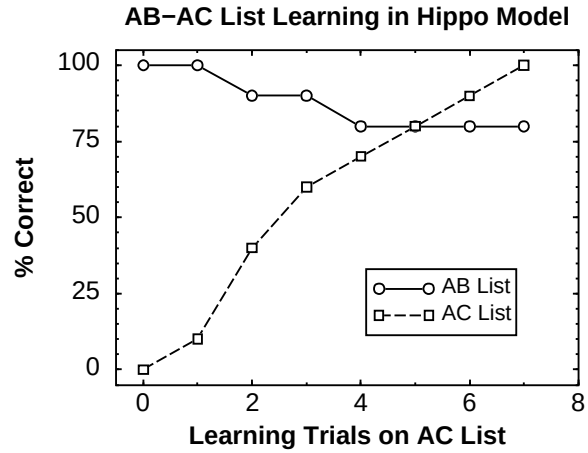

Figure 10: Results in the hippocampal model of training on the AC list after training on the AB list. Testing performance on the AB items decreases due to interference as the AC list is learned, but this interference is by no means catastrophic.

patterns as in the cortical network, and the list context was similarly a slightly perturbed version of two different list-prototype patterns. During testing, the second associate (either B or C) was omitted, requiring pattern completion in the hippocampus to fill it in (in the EC_out layer) based on the partial cue of the A stimulus and the context.

Each epoch of training consists of the 10 list items, either from the AB or AC list. There were 5 initial epochs of training to learn the AB list to 100% accuracy, and then 5 more epochs to subsequently learn the AC list to 100% accuracy. The network's performance can be measured in terms of how much of the 2nd associate (B or C) is produced in response to an input of the first associate (A) and context during testing. O'Reilly and Munakata (2000) measured this in terms of two variables: `stim_err_on` and `stim_err_off`. `stim_err_on` measures the proportion of units that were erroneously activated in `EC_out` (i.e., active but not present in the target associate pattern), and `stim_err_off` measures the proportion of units that were erroneously *not* activated in `EC_out` (i.e., not active in the network's response, but present in the target pattern). When both of these measures are near zero, then the network has correctly recalled the target associate pattern. A large `stim_err_on` indicates that the network has *confabulated* or otherwise recalled a different pattern than the cued one. This is relatively rare in the model (O'Reilly et al., 1998). A large `stim_err_off` indicates that the network has failed to recall much of the probe pattern. This is common, especially for untrained patterns.

Figure 10 shows a plot like those shown previously

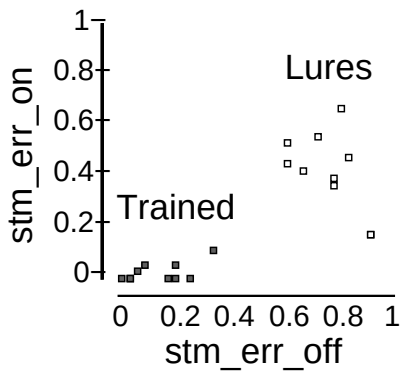

Figure 11: Composite graph log display for testing the hippocampal network, showing two kinds of errors — units that should be off but were erroneously on (`stm_err_on`), and units that should be on but were erroneously off (`stm_err_off`). The training items have relatively few of both types of errors, and the lure items are mostly inactive (high `stm_err_off`.

(figures 1 and 8), graphing testing performance on the originally-learned AB list items as the network learns the AC list. As this figure makes clear, this hippocampal network model can rapidly learn new information without suffering extensive amounts of interference on previously-learned information.

If you were to observe the network as it is tested on the testing items (missing the second associate), you would see the EC_in layer initially activated by the 1st associate and the list context. Then, this activation flows up to the CA3 layer, where pattern completion via recurrent collaterals and feedforward projections results in the activation of the original, complete CA3 representation for this item as developed during training. This completed CA3 representation then activates the CA1 representation via learned associations, and the CA1 can then fill in the missing parts of the EC_out pattern. In addition to the "inner loop" of CA3 pattern completion, there is an "outer loop" of pattern completion that occurs through partial EC_out activation feeding back into EC_in and then back up through the hippocampal system again.

We can obtain additional insight into the hippocampus by observing the model's responses to novel (unstudied) items. To do this, we also tested the network on a set of novel *lure* items in addition to the AB and AC list items. Figure 11 shows testing results on AB items and lure items, with the `stim_err_on` plotted on the Y axis against `stim_err_off` on the X axis. Each event shows up as a dot at a particular location. To the extent that these dots are in the lower left hand corner, the network is recalling accurately. You can see that the trained items produce low amounts of both types of error, and the lures produce high amounts of off-errors (i.e., they simply fail to activate units). This very sharp separation between trained items and lures has important implications for the performance of the hippocampus in memory tests, and is substantiated by analyses of behavioral data (see O'Reilly et al., 1998 for more discussion).

Summary and Discussion

The O'Reilly and Munakata (2000) model demonstrates that the unique biological properties of the hippocampus can lead to rapid learning of list items while minimizing interference effects. Because the hippocampal system employs sparse representations, it encodes the list items with highly conjunctive, pattern separated representations. These representations are the key to avoiding interference, as they use different sets of weights to encode items on different lists. In addition to pattern separation, this model demonstrates the crucial, complementary function of pattern completion, which is required to retrieve the previously-learned paired associates. We also observed that the network exhibits a very clear separation between trained and novel items, which can be contrasted with a more graded memory signal, as we discuss in a later section. In short, this network behaves much more like the human episodic memory system than the simple cortical network explored previously. This performance supports the idea that the brain needed to develop two complementary learning systems to satisfy the complementary learning objectives of rapid learning of separate events and slow integration across episodes to extract general statistics.

Other Applications of the Principles

The principles of complementary cortical and hippocampal learning mechanisms have been applied to a number of different domains, as briefly summarized in the following sections. In most cases, the same neural network model as we just described was used to simulate the empirical data, providing a compelling demonstration that the principles are sufficient to account for a wide range of findings.

Rapid Incidental Conjunctive Learning Tasks

The principles we have outlined clearly suggest that the hippocampus should be most important for tasks that involve the rapid learning of conjunctive information, as is characteristic of human episodic memory. When we encode the events of our daily lives, we do so without expending deliberate effort, and we do so rapidly because these events are by definition fleeting in nature — we need to encode them as they happen. The are also conjunctive in nature because they bind together all the

many elements of an event into a unitary representation that says, "all these things were present at the same time" (e.g., a particular room, furniture, collection of people, actions, etc).

Fortunately, a number of tasks that capture the *rapid, incidental conjunctive learning* characteristics of the hippocampus have been recently developed in animals. With the use of selective hippocampal lesions in these tasks, experimenters have confirmed that the hippocampus is critical for this function. In these tasks, subjects are exposed to a set of features in a particular configuration, and then the features are rearranged. Subjects are then tested to determine if they can detect the rearrangement. If the test indicates that the rearrangement was detected, then one can infer the subject learned a conjunctive representation of the original configuration.

Perhaps the simplest demonstration comes from the study of the role of the hippocampal formation in exploratory behavior. Control rats and rats with damage to the dorsal hippocampus were repeatedly exposed to a set of objects that were arranged on a circular platform in a fixed configuration relative to a large and distinct visual cue (Save, Poucet, Foreman, & Buhot, 1992). After the exploratory behavior of both sets of rats habituated, the same objects were rearranged into a different configuration. This rearrangement reinstated exploratory behavior in the control rats but not in the rats with damage to the hippocampus. In a third phase of the study, a new object was introduced into the mix. This manipulation reinstated exploratory behavior in both sets of rats. This pattern of data suggests that both control rats and rats with damage to the hippocampus encode representations of the individual objects and can discriminate them from novel objects. However, only the control rats encoded the conjunctions necessary to represent the spatial arrangement of the objects, even though this was not in any way a requirement of the task. Several other studies of this general form have found similar results in rats (Honey, Watt, & Good, 1998; Honey & Good, 1993; Good & Bannerman, 1997; Hall & Honey, 1990; Honey, Willis, & Hall, 1990). In humans, the well established incidental context effects on memory (e.g., Godden & Baddeley, 1975) have been shown to be hippocampal-dependent (Mayes, MacDonald, Donlan, & Pears, 1992). Other hippocampal incidental conjunctive learning effects have also been demonstrated in humans (Chun & Phelps, 1999).

We have shown that the same neural network model constructed according to our principles and tested on the AB-AC task as described above exhibits a clear hippocampal sensitivity in these rapid incidental conjunctive learning tasks (O'Reilly & Rudy, 2001). By extension, we therefore believe that the model accounts for the involvement of the hippocampus in episodic memory in humans.

Contextual Fear Conditioning

Evidence for the involvement of the hippocampal formation in the incidental learning of stimulus conjunctions has also emerged in the contextual fear conditioning literature. This example also provides a simple example of the widely-discussed role of the hippocampus in spatial learning (e.g., O'Keefe & Nadel, 1978; McNaughton & Nadel, 1990). Rats with damage to the hippocampal formation do not express fear to a context or place where shock occurred, but will express fear to an explicit cue (e.g., a tone) paired with shock (Kim & Fanselow, 1992; Phillips & LeDoux, 1994; but see Maren, Aharonov, & Fanselow, 1997). Rudy and O'Reilly (1999) recently provided specific evidence that, in intact rats, the context representations are conjunctive in nature, which has been widely assumed (e.g., Fanselow, 1990; Kiernan & Westbrook, 1993; Rudy & Sutherland, 1994). For example, we compared the effects of preexposure to the conditioning context with the effects of preexposure to the separate features that made up the context. Only preexposure to the intact context facilitated contextual fear conditioning, suggesting that conjunctive representations across the context features were necessary. We also showed that pattern completion of hippocampal conjunctive representations can lead to generalized fear conditioning. Furthermore, a recent study showed that rats can condition to a *memory* of a context, while they are actually located in a novel environment (Rudy & O'Reilly, 2001). The memory is activated via pattern completion from a bucket that was reliably associated with the context during preexposure.

We have simulated the incidental learning of conjunctive context representations in fear conditioning using the same principles as described above (O'Reilly & Rudy, 2001). For example, Figure 12 shows the rat and model data for the separate versus intact context features experiment from Rudy and O'Reilly (1999), with the model providing a specific prediction regarding the effects of hippocampal lesions, which has yet to be tested empirically.

Conjunctions and Nonlinear Discrimination Learning

One important application of the conjunctive representations idea has been to *nonlinear discrimination problems*. These problems require conjunctive representations to solve because each of the individual stimuli is ambiguous (equally often rewarded and not rewarded). The negative patterning problem is a good example. It

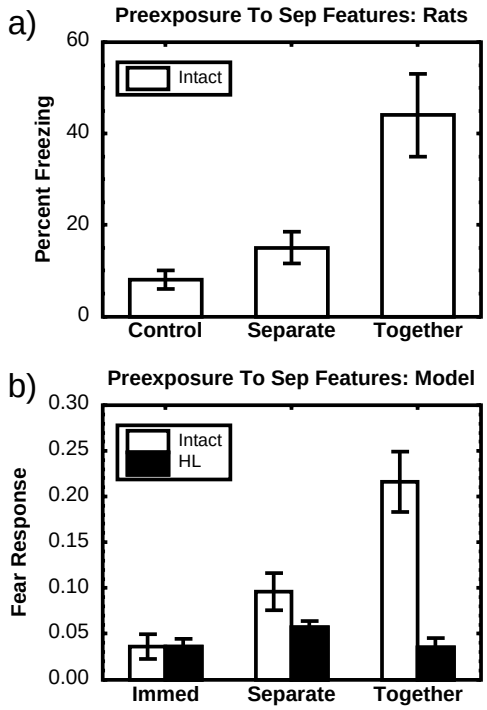

Figure 12: Effects of exposure to the features separately compared to exposure to the entire context on level of fear response in **a)** rats (data from Rudy and O'Reilly (1999)) and the model (O'Reilly and Rudy (2001)). The immediate shock condition (Immed) is included as a control condition for the model. Intact rats and the intact model show a significant effect of being exposed to the entire context together compared to the features separately, while the hippocampally lesioned model exhibits slightly more responding in the separate condition, possibly because of the greater overall number of training trials in this case.

involves two stimuli, A and B (e.g., a light and a tone), which are associated with reward (indicated by $+$) or not ($-$). Three different trial types are trained: $A+$, $B+$, $AB-$. Thus, the conjunction of the two stimuli ($AB-$) must be treated differently from the two stimuli separately ($A+$, $B+$). A conjunctive representation that forms a novel encoding of the two stimuli together can facilitate this form of learning. Therefore, the fact that hippocampal damage impairs learning the negative patterning problem (Alvarado & Rudy, 1995; Rudy & Sutherland, 1995; McDonald, Murphy, Guarraci, Gortler, White, & Baker, 1997) would appear to support the idea that the hippocampus employs pattern separated, conjunctive representations. However, it is now clear that a number of other nonlinear discrimination learning problems are unimpaired by hippocampal damage (Rudy & Sutherland, 1995).

The general explanation of these results according to the full set of principles outlined above is that:

- The explicit task demands present in a nonlinear discrimination learning problem cause the cortex alone (with a lesioned hippocampus) to learn the task via error-driven learning.
- Nonlinear discrimination problems take many trials to learn even in intact animals, allowing the slow cortical learning to accumulate a solution.
- The absence of hippocampal learning speed advantages in normal rats, despite the more rapid hippocampal learning rate, can be explained by the fact that the hippocampus is engaging in pattern completion in these problems, instead of pattern separation.

We have substantiated this verbal account by running computational neural network simulations that embodied the principles developed above (O'Reilly & Rudy, 2001). These simulations showed that in many — but not all — cases, removing the hippocampal component did not significantly impair learning performance on nonlinear discrimination learning problems, matching the empirical data. To summarize, this work showed that it is essential to go beyond a simple conjunctive story and include a more complete set of principles in understanding hippocampal and cortical function. Because this more complete set of principles, implemented in an explicit computational model, accounts for the empirical data, this data provides support for these principles.

Dual-Process Memory Models

The dual mechanisms of neocortex and hippocampus provide a natural fit with dual-process models of recognition memory (Jacoby, Yonelinas, & Jennings, 1997; Aggleton & Shaw, 1996; Aggleton & Brown, 1999; Vargha-Khadem, Gadian, Watkins, Connelly, Van Paesschen, & Mishkin, 1997; Holdstock, Mayes, Roberts, Cezayirli, Isaac, O'Reilly, & Norman, 2002; O'Reilly et al., 1998; Norman & O'Reilly, in press). These models hold that recognition can be subserved by two different processes, a *recollection* process and a *familiarity* process. Recollection involves the recall of specific episodic details about the item, and thus fits well with the hippocampal principles developed here. Indeed, we have simulated distinctive aspects of recollection using essentially the same model (Norman & O'Reilly, in press; O'Reilly et al., 1998). Familiarity is a non-specific sense that the item has been seen recently — we argue that this can be subserved by the small weight changes produced by slow cortical learning. Current simulation work has shown that a simple cortical model can account for a number of distinctive properties of the familiarity signal (Norman & O'Reilly, in press).

One specific and somewhat counter-intuitive prediction of our principles has recently been confirmed

empirically in experiments on a patient with selective hippocampal damage (Holdstock et al., 2002). This patient showed intact recognition memory for studied items compared to similar lures when tested in a two-alternative forced-choice procedure (2AFC), but was significantly impaired relative to controls for the same kinds of stimuli using a single item yes-no (YN) procedure. We argue that because the cortex uses overlapping distributed representations, the strong similarity of the lures to the studied items produces a strong familiarity signal for these lures (as a function of this overlap). When tested in a YN procedure, this strong familiarity of the lures produces a large number of false alarms, as was observed in the patient. However, because the studied item has a small but reliably stronger familiarity signal than the similar lure, this strength difference can be detected in the 2AFC version, resulting in normal recognition performance in this condition. The normal controls, in contrast, have an intact hippocampus which performs pattern separation and is able to distinguish the studied items and similar lures, regardless of the testing format.

Comparison with Other Approaches

A number of other approaches to understanding cortical and hippocampal function share important similarities with our approach, including for example the use of Hebbian learning and pattern separation (e.g., Hasselmo, 1995; McNaughton & Nadel, 1990; Touretzky & Redish, 1996; Burgess & O'Keefe, 1996; Wu, Baxter, & Levy, 1996; Treves & Rolls, 1994; Moll & Miikkulainen, 1997; Alvarez & Squire, 1994). These other approaches all offer other important principles, many of which would be complementary to those discussed here so that it would be possible to add them to a larger, more complete model.

Perhaps the largest area of disagreement is in terms of the relative independence of the cortical learning mechanisms from the hippocampus. There are several computationally-explicit models that propose the neocortex is incapable of powerful learning without the help of the hippocampus (Gluck & Myers, 1993; Schmajuk & DiCarlo, 1992; Rolls, 1990), and other more general theoretical views that express a similar notion of limited cortical learning with hippocampal damage (Glisky, Schacter, & Tulving, 1986; Squire, 1992; Cohen & Eichenbaum, 1993; Wickelgren, 1979; Sutherland & Rudy, 1989). In contrast, our principles hold that the cortex alone is a highly capable learning system, that can for example learn complex conjunctive representations in the service of nonlinear discrimination learning problems. We think the growing literature on preserved learning with focal hippocampal damage supports the idea that the cortex by itself is a powerful learning system.

Summary

We have shown that a small set of computationally-motivated principles can account for a wide range of empirical findings regarding the differential properties of the neocortex and hippocampus in learning and memory. These principles go beyond accounting for data by providing clear reasons why the brain has the specialized areas that it does, and what the mechanistic differences are between these areas.

References

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

Aggleton, J. P., & Shaw, C. (1996). Amnesia and recognition memory: A re-analysis of psychometric data. *Neuropsychologia*, *34*, 51–62.

Alvarado, M. C., & Rudy, J. W. (1995). A comparison of kainic acid plus colchicine and ibotenic acid induced hippocampal formation damage on four configural tasks in rats. *Behavioral Neuroscience*, *109*, 1052–1062.

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.

Barnes, C. A., McNaughton, B. L., Mizumori, S. J. Y., Leonard, B. W., & Lin, L.-H. (1990). Comparison of spatial and temporal characteristics of neuronal activity in sequential stages of hippocampal processing. *Progress in Brain Research*, *83*, 287–300.

Barnes, J. M., & Underwood, B. J. (1959). Fate of first-list associations in transfer theory. *Journal of Experimental Psychology*, *58*, 97–105.

Boss, B. D., Peterson, G. M., & Cowan, W. M. (1985). On the numbers of neurons in the dentate gyrus of the rat. *Brain Research*, *338*, 144–150.

Boss, B. D., Turlejski, K., Stanfield, B. B., & Cowan, W. M. (1987). On the numbers of neurons in fields CA1 and CA3 of the hippocampus of Sprague-Dawley and Wistar rats. *Brain Research*, *406*, 280–287.

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

Chun, M. M., & Phelps, E. A. (1999). Memory deficits for implicit contextual information in amnesic subjects with hippocampal damage. *Nature Neuroscience*, *2*(9), 844–847.

Cohen, N. J., & Eichenbaum, H. (1993). *Memory, amnesia, and the hippocampal system*. Cambridge, MA: MIT Press.

Collingridge, G. L., & Bliss, T. V. P. (1987). NMDA receptors — their role in long-term potentiation. *Trends in Neurosciences*, *10*, 288–293.

Crick, F. H. C. (1989). The recent excitement about neural networks. *Nature*, *337*, 129–132.

Fanselow, M. S. (1990). Factors governing one-trial contextual conditioning. *Animal Learning and Behavior*, *18*, 264–270.

French, R. M. (1992). Semi-distributed representations and catastrophic forgetting in connectionist networks. *Connection Science*, *4*, 365–377.

Glisky, E. L., Schacter, D. L., & Tulving, E. (1986). Computer learning by memory-impaired patients: Acquisition and retention of complex knowledge. *Neuropsychologia*, *24*, 313–328.

Gluck, M. A., & Myers, C. E. (1993). Hippocampal mediation of stimulus representation: A computational theory. *Hippocampus*, *3*, 491–516.

Godden, D. R., & Baddeley, A. D. (1975). Context-dependent memory in two natural environments: On land and under water. *British Journal of Psychology*, *66*, 325–331.

Good, M., & Bannerman, D. (1997). Differential effects of ibotenic acid lesions of the hippocampus and blockade of n-methyl-d-aspartate receptor-dependent long-term potentiation on contextual processing in rats. *Behavioral Neuroscience*, *111*, 1171–1183.

Hall, G., & Honey, R. C. (1990). Context-specific conditioning in the conditioned-emotional-response procedure. *Journal of Experimental Psychology: Animal Behavior Processes*, *16*, 271–278.

Hasselmo, M. E. (1995). Neuromodulation and cortical function: Modeling the physiological basis of behavior. *Behavioural Brain Research*, *67*, 1–27.

Hasselmo, M. E., & Wyble, B. (1997). Free recall and recognition in a network model of the hippocampus: Simulating effects of scopolamine on human memory function. *Behavioural Brain Research*, *89*, 1–34.

Hebb, D. O. (1949). *The organization of behavior*. New York: Wiley.

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.

Honey, R. C., & Good, M. (1993). Selective hippocampal lesions abolish the contextual specificity of latent inhibition and conditioning. *Behavioral Neuroscience*, *107*, 23–33.

Honey, R. C., Watt, A., & Good, M. (1998). Hippocampal lesions disrupt an associative mismatch process. *Journal of Neuroscience*, *18*, 2226–2230.

Honey, R. C., Willis, A., & Hall, G. (1990). Context specificity in pigeon autoshaping. *Learning and Motivation*, *21*, 125–136.

Jacoby, L. L., Yonelinas, A. P., & Jennings, J. M. (1997). The relation between conscious and unconscious (automatic) influences: A declaration of independence.

In J. D. Cohen, & J. W. Schooler (Eds.), *Scientific approaches to consciousness* (pp. 13–47). Mahway, NJ: Lawrence Erlbaum Associates.

Kiernan, M. J., & Westbrook, R. F. (1993). Effects of exposure to a to-be-shocked environment upon the rat's freezing response: Evidence for facilitation, latent inhibition, and perceptual learning. *Quarterly Journal of Psychology*, *46B*, 271–288.

Kim, J. J., & Fanselow, M. S. (1992). Modality-specific retrograde amnesia of fear. *Science*, *256*, 675–677.

Kortge, C. A. (1993). Episodic memory in connectionist networks. *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society* (pp. 764–771). Hillsdale, NJ: Erlbaum.

Maren, S., Aharonov, G., & Fanselow, M. S. (1997). Neurotoxic lesions of the dorsal hippocampus and Pavlovian fear conditioning. *Behavioural Brain Research*, *88*, 261–274.

Marr, D. (1969). A theory of cerebellar cortex. *Journal of Physiology (London)*, *202*, 437–470.

Marr, D. (1970). A theory for cerebral neocortex. *Proceedings of the Royal Society (London) B*, *176*, 161–234.

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

Mayes, A. R., MacDonald, C., Donlan, L., & Pears, J. (1992). Amnesics have a disproportionately severe memory deficit for interactive context. *Quarterly Journal of Experimental Psychology*, *45A*, 265–297.

McClelland, J. L., & Goddard, N. H. (1996). Considerations arising from a complementary learning systems perspective on hippocampus and neocortex. *Hippocampus*, *6*, 654–665.

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.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.), *The psychology of learning and motivation, vol. 24* (pp. 109–164). San Diego, CA: Academic Press.

McDonald, R. J., Murphy, R. A., Guarraci, F. A., Gortler, J. R., White, N. M., & Baker, A. G. (1997). Systematic comparison of the effects of hippocampal and fornix-fimbria lesions on the acquisition of three configural discriminations. *Hippocampus*, *7*, 371–388.

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.

McNaughton, B. L., & Nadel, L. (1990). Hebb-Marr networks and the neurobiological representation of action in space. In M. A. Gluck, & D. E. Rumelhart (Eds.), *Neuroscience and connectionist theory* (Chap. 1, pp. 1–63). Hillsdale, NJ: Erlbaum.

McRae, K., & Hetherington, P. A. (1993). Catastrophic interference is eliminated in pretrained networks. *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* (pp. 723–728). Hillsdale, NJ: Erlbaum.

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

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

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

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. (2001). Generalization in interactive networks: The benefits of inhibitory competition and Hebbian learning. *Neural Computation*, *13*, 1199–1242.

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. (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.

Phillips, R. G., & LeDoux, J. E. (1994). Lesions of the dorsal hippocampal formation interfere with background but not foreground contextual fear conditioning. *Learning and Memory*, *1*, 34–44.

Rolls, E. T. (1990). Principles underlying the representation and storage of information in neuronal networks in the primate hippocampus and cerebral cortex. In S. F. Zornetzer, J. L. Davis, & C. Lau (Eds.), *An introduction to neural and electronic networks* (pp. 73–90). San Diego, CA: Academic Press.

Rudy, J. W., & O'Reilly, R. C. (1999). Contextual fear conditioning, conjunctive representations, pattern completion, and the hippocampus. *Behavioral Neuroscience*, *113*, 867–880.

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

Rudy, J. W., & Sutherland, R. J. (1994). The memory coherence problem, configural associations, and the hippocampal system. In D. L. Schacter, & E. Tulving (Eds.), *Memory systems 1994* (pp. 119–146). Cambridge, MA: MIT Press.

Rudy, J. W., & Sutherland, R. W. (1995). Configural association theory and the hippocampal formation: An appraisal and reconfiguration. *Hippocampus*, *5*, 375–389.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533–536.

Save, E., Poucet, B., Foreman, N., & Buhot, N. (1992). Object exploration and reactions to spatial and nonspatial changes in hooded rats following damage to parietal cortex or hippocampal formation. *Behavioral Neuroscience*, *106*, 447–456.

Schmajuk, N. A., & DiCarlo, J. J. (1992). Stimulus configuration, classical conditioning, and hippocampal function. *Psychological Review*, *99*(2), 268–305.

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

Sloman, S. A., & Rumelhart, D. E. (1992). Reducing interference in distributed memories through episodic gating. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *Essays in honor of W. K. Estes* (pp. 227–248). Hillsdale, NJ: Erlbaum.

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

Squire, L. R., Shimamura, A. P., & Amaral, D. G. (1989). Memory and the hippocampus. In J. H. Byrne, & W. O. Berry (Eds.), *Neural models of plasticity: Experimental and theoretical approaches* (pp. 208–239). San Diego, CA: Academic Press.

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.

Tamamaki, N. (1991). The organization of reciprocal connections between the subiculum, field CA1 and the entorhinal cortex in the rat. *Society for Neuroscience Abstracts*, *17*, 134.

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.

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.

White, H. (1989). Learning in artificial neural networks: A statistical perspective. *Neural Computation*, *1*, 425–464.

Wickelgren, W. A. (1979). Chunking and consolidation: A theoretical synthesis of semantic networks, configuring in conditioning, S-R versus cognitive learning, normal forgetting, the amnesic syndrome, and the hippocampal arousal system. *Psychological Review*, *86*, 44–60.

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.

Zipser, D., & Andersen, R. A. (1988). A backpropagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature*, *331*, 679–684.