# Computational Principles of Learning in the Neocortex and Hippocampus

Randall C. O'Reilly & Jerry W. Rudy

Department of Psychology
University of Colorado Boulder
Campus Box 345
Boulder, CO 80309
oreilly@psych.colorado.edu

11th October 2000

Published in *Hippocampus*, 2000, vol 10, p. 389-397

## Abstract

We present an overview of our 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. We summarize the results of applying these principles to a wide range of phenomena in conditioning, habituation, contextual learning, recognition memory, recall, and retrograde amnesia, and point to directions of current development.

## Introduction

This paper presents a computational approach towards understanding the different contributions of the neocortex and hippocampus in learning and memory. This approach uses basic principles of computational neural network learning mechanisms to understand both *what* is different about the way these two neural systems learn, and *why* they should have these differences. Thus, the computational approach can go beyond mere description towards understanding the deeper principles underlying the organization of the cognitive system. These principles are based on an convergence of biological, psychological, and computational constraints, and serve to bridge between these different levels of analysis.

The set of principles discussed in this paper 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, 1999, in press). 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. This paper provides a concise summary of the previous work, and a discussion of current and future directions.

## The Principles

There are several levels of principles that can be distinguished by their degree of specificity in characterizing the nature of the underlying mechanisms. We begin with the most basic principles and proceed towards greater specificity.

### *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).

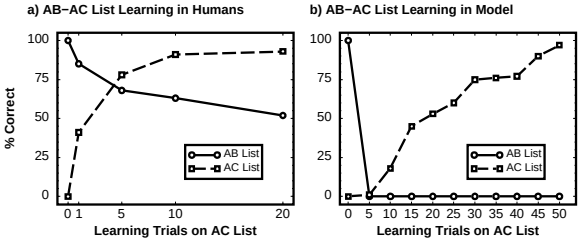


Figure 1: Human and model data for AB-AC list learning. a) Humans show some interference for the AB list items as a function of new learning on the AC list items. b) Model shows a catastrophic level of interference. (data reproduced from McCloskey & Cohen, 1989).

The first set of principles concern the effects of overlapping representations (i.e., shared units between two different distributed representations) and rate of learning on the ability to rapidly learn new information with a level of interference characteristic of human subjects:

- 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 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.

### *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). Furthermore, overlapping representations facilitate this integration process, because the

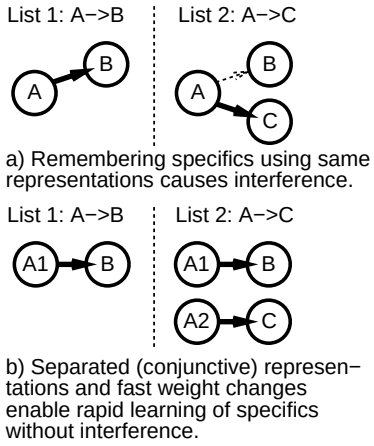


Figure 2: Interference as a function of overlapping (same) representations versus separated representations. **a)** Using the same representation to encode two different associations ($A \rightarrow B$ and $A \rightarrow C$) causes interference — the subsequent learning of $A \rightarrow C$ interferes 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.

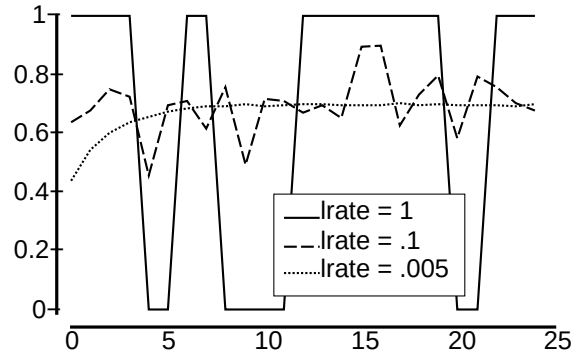


Figure 3: Weight value learning about a single input unit that is either active or not. The weight increases when the input is on, and decrease when it is off, in proportion to the size of the learning rate. The input has an overall probability of being active of .7. Larger learning rates (.1 or 1) lead 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.

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 all neural networks are 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. This line of reasoning provides a strikingly good fit to the known properties of the hippocampus and neocortex, respectively. The details of this fit in various contexts is the substance of the remainder of the paper, but the general idea is that:

- 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). Before discussing the details, a few more principles need to be developed first.

## *Conjunctive Representations and Nonlinear Discrimination Learning*

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, 1999). 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, 1999). To summarize:

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

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 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 next set of principles help to make sense of these data so that they can be reconciled with our interference-based principles of conjunctive pattern separation, as discussed in a subsequent section.

## *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 — this can complicate the involvement of the hippocampus in nonlinear discrimination problems.

### *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, 1999; 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, 1999):

**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 section, we highlight the application of the principles presented here to learning and memory phenomena involving both the cortex and hippocampus.

## Applications of the Principles

The principles just developed have been applied to a number of different domains, as summarized in the following sections. In most cases, the same neural network model developed according to these principles (O'Reilly & Rudy, 1999) was used to simulate the empirical data, providing a compelling demonstration that the principles are sufficient to account for a wide range of findings.

### *Conjunctions and Nonlinear Discrimination Learning*

We first apply the above principles to the puzzling pattern of hippocampal involvement in nonlinear discrimination learning problems. The general statement of the issue is that although we think the hippocampus is specialized for encoding conjunctive bindings of stimuli (and keeping these separated from each other to minimize interference), apparently direct tests of this idea in the form of nonlinear discrimination learning problems have not provided clear support. Specifically, rats with hippocampal lesions can learn a number of these nonlinear discrimination problems just like intact rats. 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 substantiated this verbal account by running computational neural network simulations that embodied the principles developed above (O'Reilly & Rudy, 1999; Figure 4). 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. Figure 5 shows one specific example, where the negative patterning problem discussed previously ($A+$, $B+$, $AB-$) is impaired with hippocampal lesions, but performance is not impaired on a very similar *ambiguous feature* problem, $AC+$, $B+$, $AB-$, $C-$,

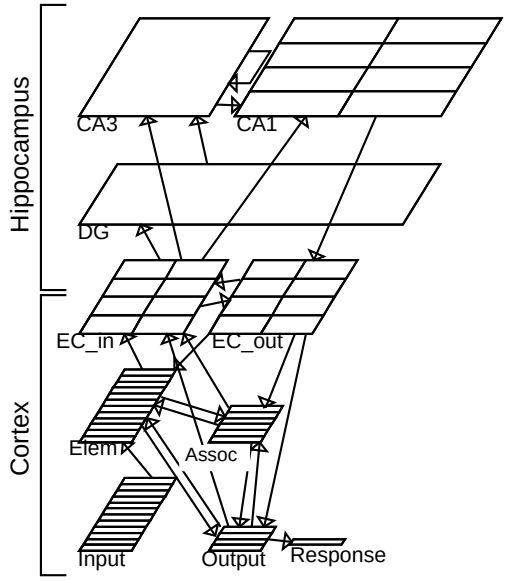


Figure 4: The O'Reilly and Rudy (1999) model, showing both cortical and hippocampal components. The cortex has 12 different input dimensions (sensory pathways), with 4 different values per dimension. These are represented separately in the elemental cortex (Elem). Higher level association cortex (Assoc) can form conjunctive representations of these elements, if demanded by the task. The interface to the hippocampus is via the entorhinal cortex, which contains a one-to-one mapping of the elemental, association, and output cortical representations. The hippocampus can reinstate a pattern of activity over the cortex via the EC.

(Gallagher & Holland, 1992). See O'Reilly and Rudy (1999) for a detailed discussion of the differential performance on these tasks.

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.

### *Rapid Incidental Conjunctive Learning Tasks*

A consideration of the full set of principles suggests that another class of tasks might provide a much better measure of hippocampal learning compared to the nonlinear discrimination problems suggested by Sutherland and Rudy (1989). As we just saw, the very fact that nonlinear discrimination problems *require* conjunctive representations is what drives the cortex alone to be able to solve them via error-driven learning. Therefore, we suggest that *incidental* conjunctive learning tasks, where conjunctive representations are not forced by specific

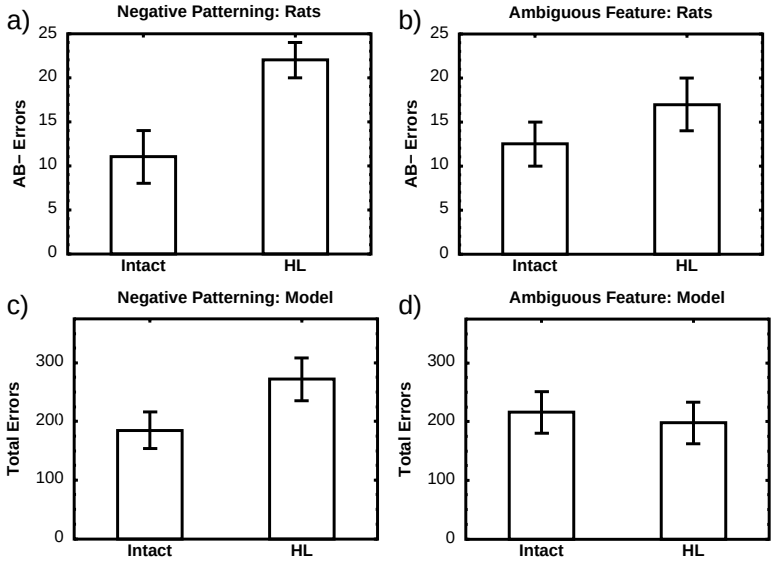


Figure 5: Results for the negative patterning (left column) and ambiguous feature ($AC+$, $B+$, $AB-$, $C-$; right column) problems. The top row shows data from rats from Alvarado and Rudy (1995), and the bottom row shows data from the model. *Intact* is intact rats/networks, and *HL* is rats/networks with hippocampal lesions. N=40 different random initializations for the model. The hippocampally-lesioned system is able to learn the problems, and all conditions require many trials (i.e., large number of errors). Negative patterning is differentially impaired with a hippocampal lesion. Data from O'Reilly and Rudy (1999).

task demands, may provide a much better index of hippocampal function (O'Reilly & Rudy, 1999). Furthermore, the task should only allow for a relatively brief period of learning, which will emphasize the rapid learning of the hippocampus as compared to the slow learning of the cortex. Thus, we characterize these tasks as *rapid, incidental conjunctive learning tasks*.

There are several recent studies of tasks that fit the rapid, incidental conjunctive characterizations. 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. The literature indicates that the incidental learning of stimulus conjunctions, unlike many nonlinear discrimination problems, *is* dependent on the hippocampus.

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 nonlinear discrimination learning problems as described above exhibits a clear hippocampal sensitivity in these rapid incidental conjunctive learning tasks (O'Reilly & Rudy, 1999).

## *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.

We have simulated the incidental learning of conjunctive context representations in fear conditioning using the same principles as described above (O'Reilly & Rudy, 1999). For example, Figure 6 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.

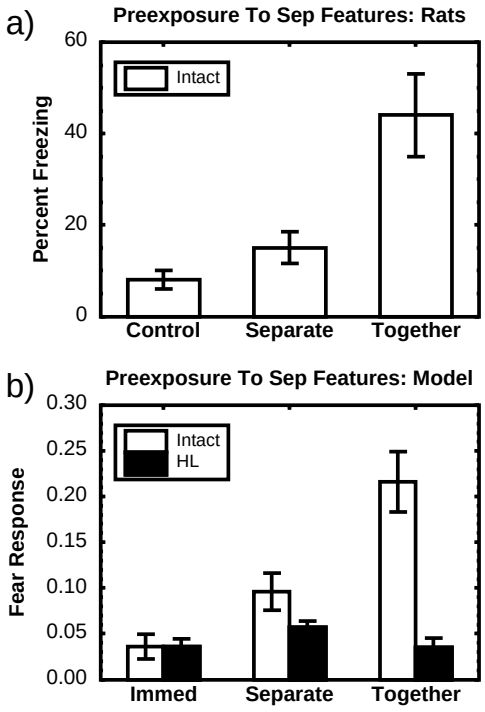


Figure 6: Effects of exposure to the features separately compared to exposure to the entire context on level of fear response in **a)** rats (data from Rudy and O'Reilly (1999)) and the model. The immediate shock condition (Immed) is included as a control condition for the model. Intact rats and the 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. Simulation data from O'Reilly and Rudy (1999).

## *Transitivity and Flexibility*

Whereas the previous examples concern the learning of conjunctive representations, this next example is concerned with the flexible use of learned information. Several theorists have described memories encoded by the hippocampus as being flexible, meaning that (a) such memories can be applied inferentially in novel situations (Eichenbaum, 1992; O'Keefe & Nadel, 1978) or (b) that they are available to multiple response systems (Squire, 1992). Although the term *flexibility* provides a useful description of certain behaviors, it does not provide a mechanistic understanding of how this flexibility arises from the properties of the hippocampus.

We have shown that hippocampal pattern completion plays an important role in producing this flexible behavior (O'Reilly & Rudy, 1999). Specifically, we showed that the *transitivity* studies of Bunsey and Eichenbaum (1996) and Dusek and Eichenbaum (1997) can be simulated using the same model as in all of the previous examples, with pattern completion playing a key role. Interestingly, the model shows that the training parameters employed in these studies interact significantly with the pattern completion mechanism to produce the observed transitivity effects. We are able to make a number of novel empirical predictions that are inconsistent with a simple logical-reasoning mechanism by manipulating these factors (O'Reilly & Rudy, 1999).

## *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, in press; O'Reilly et al., 1998). These models hold that recognition can be subserved by two different processes, a *recollection* process and a *familiarity* process. Recollection involves the recall of specific episodic details about the item, and thus fits well with the hippocampal principles developed here. Indeed, we have simulated distinctive aspects of recollection using essentially the same model (O'Reilly et al., 1998). Familiarity is a non-specific sense that the item has been seen recently — we argue that this can be subserved by the small weight changes produced by slow cortical learning. Current simulation work has shown that a simple cortical model can account for a number of distinctive properties of the familiarity signal (Norman, O'Reilly, & Huber, 2000).

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., in press). 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.

### *Retrograde Amnesia*

Several lines of empirical evidence suggest that there is a *retrograde gradient* for memory loss as a function of hippocampal damage, with the most *recent* memories being the most severely affected, while older memories are relatively intact (e.g., Squire, 1992; Winocur, 1990; Kim & Fanselow, 1992; Zola-Morgan & Squire, 1990). Theoretically, this phenomenon can be understood in terms of the cortex gradually acquiring hippocampal information (e.g., McClelland et al., 1995; Alvarez & Squire, 1994). However, this account has been called into question recently, both from failures to replicate the retrograde findings (Sutherland, Weisend, Mumby, Astur, Hanlon, Koerner, & Thomas, in press), and reinterpretations of the existing findings in ways that do not require that the cortex acquires information from the hippocampus (e.g., the "multiple hippocampal trace" theory of Nadel & Moscovitch, 1997).

Our computational principles suggest that to the extent there are opportunities for the hippocampus to reactivate cortical patterns of activity, the consequent cortical learning will necessarily produce a consolidation-like effect. We were able to fit a number of different retrograde amnesia gradients using these principles (McClelland et al., 1995).

## 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, 1997; 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.

Empirically, the data that appears to support the limited cortical learning view tends to be based on larger lesions of the medial temporal lobe. With the advent of selective lesion techniques in rats and monkeys, and the study of people with highly selective hippocampal lesions, it is becoming clear that the cortex is capable of quite substantial learning on its own. Perhaps the most dramatic evidence comes from a group of human amnesics who suffered bilateral selective hippocampal damage at relatively young ages (Vargha-Khadem et al., 1997). Despite having significantly impaired hippocampal function (as was supported by brain scans and very

poor performance on recall tests), these individuals had acquired normal or nearly normal levels of cognitive functioning in language, semantic knowledge, and had normal or nearly-normal IQs. Although it is difficult to completely rule out the idea that this preserved semantic learning is the result of residual hippocampal functioning (as advocated by Squire & Zola, 1998), this seems somewhat implausible in the face of the patient's significant recall impairments and the brain scan evidence.

One important conclusion from this line of reasoning is that the cortical regions surrounding the hippocampus in the medial temporal lobes are particularly important for many kinds of learning and memory. We suggest that this is because of a significant convergence of other cortical association areas in these regions (O'Reilly & Rudy, 1999; Mishkin, Suzuki, Gadian, & Vargha-Khadem, 1997; Mishkin, Vargha-Khadem, & Gadian, 1998).

## 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. In addition, these principles make a large number of empirical predictions that will be tested in future research.

## 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.

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

Bunsey, M., & Eichenbaum, H. (1996). Conservation of hippocampal memory function in rats and humans. *Nature, 379*, 255.

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.

Dusek, J. A., & Eichenbaum, H. (1997). The hippocampus and memory for orderly stimulus relations. *Proceedings of the National Academy of Sciences, 94*, 7109–7114.

Eichenbaum, H. (1992). The hippocampal system and declarative memory in animals. *Journal of Cognitive Neuroscience, 4*(3), 217–231.

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.

Gallagher, M., & Holland, P. C. (1992). Preserved configural learning and spatial learning impairment in rats with hippocampal damage. *Hippocampus*, *2*, 81–88.

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.

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

Hasselmo, M. E. (1997). euromodulation 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*, *67*, 1–27.

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. (in press). Memory dissociations following human hippocampal damage. *Hippocampus*.

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.

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. (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., 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 connectionst 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, Inc.

McDonald, R. J., Murphy, R. A., Guarraci, F. A., Gortler, J. R., White, & Baker, A. G. (1997). Systemic 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: Lawrence Erlbaum Associates.

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.

Mishkin, M., Suzuki, W., Gadian, D. G., & Vargha-Khadem, F. (1997). Hierarchical organization of cognitive memory. *Philosophical Transactions of the Royal Society, London, B, 352*, 1461–1467.

Mishkin, M., Vargha-Khadem, F., & Gadian, D. G. (1998). Amnesia and the organization of the hippocampal system. *Hippocampus, 8*, 212–216.

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

Nadel, L., & Moscovitch, M. (1997). Memory consolidation, retrograde amnesia and the hippocampal complex. *Current Opinion in Neurobiology, 7*, 217.

Norman, K. A., O'Reilly, R. C., & Huber, D. E. (2000). Modeling neocortical contributions to recognition memory. *The Cognitive Neuroscience Meeting, 2000*.

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford: 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., & 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., & 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. (1999). *Conjunctive representations in learning and memory: Principles of cortical and hippocampal function* (Institute of Cognitive Science TR 99-01). Boulder: University of Colorado Boulder.

O'Reilly, R. C., & Rudy, J. W. (in press). Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psychological Review*.

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., & 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 non-spatial 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., & Zola, S. M. (1998). Episodic memory, semantic memory, and amnesia. *Hippocampus, 8*, 205–211.

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.

Sutherland, R. J., Weisend, M. P., Mumby, D., Astur, R. S., Hanlon, F. M., Koerner, A., & Thomas, M. J. (in press). Retrograde amnesia after hippocampal damage: Recent vs. remote memories in several tasks. *Hippocampus*.

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.

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.

Winocur, G. (1990). Anterograde and retrograde amnesia in rats with dorsal hippocampal or dorsomedial thalamic lesions. *Behavioural Brain Research, 38*, 145–154.

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.

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.

Zola-Morgan, S., & Squire, L. R. (1990). The primate hippocampal formation: Evidence for a time-limited role in memory storage. *Science, 250*, 288–290.