Institute of Cognitive Science

Technical Report

University of Colorado, Boulder

Conjunctive Representations in Learning and Memory: Principles of Cortical and Hippocampal Function

Randall C. O'Reilly and Jerry W. Rudy
oreilly@psych.colorado.edu jrudy@clipr.colorado.edu

Department of Psychology
University of Colorado
Boulder, CO 80309

Technical Report 99-01

Abstract

We present a theoretical framework for understanding the roles of the hippocampus and neocortex in learning and memory. This framework incorporates a theme found in many theories of hippocampal function, that the hippocampus is responsible for developing conjunctive representations binding together stimulus elements into a unitary representation that can later be recalled from partial input cues. This idea appears problematic, however, because it is contradicted by the fact that hippocampally lesioned rats can learn nonlinear discrimination problems that require conjunctive representations. Our framework accommodates this finding by establishing a principled division of labor between the cortex and hippocampus, where the cortex is responsible for slow learning that integrates over multiple experiences to extract generalities, while the hippocampus performs rapid learning of the arbitrary contents of individual experiences. This framework shows that nonlinear discrimination problems are not good tests of hippocampal function, and suggests that tasks involving rapid, incidental conjunctive learning are better. We implement this framework in a computational neural network model, and show that it can account for a wide range of data in animal learning, thus validating our theoretical ideas, and providing a number of insights and predictions about these learning phenomena.

Contents

Introduction

The role of the hippocampus in memory has been characterized in many different ways, but one common thread is the idea that the hippocampus binds together the sensory features of a situation or episode to create a unitary representation of the experience. Thus, the hippocampus is said to construct *configural* representations, support the acquisition of a *spatial map*, represent the *conjunction* or *co-occurrence* of the stimulus features, or to *chunk* or *bind* these features into a unitary representation. This binding process enables the the original conjunction of features to be recalled from a subset of its parts, and allows the conjunction to be treated differently from the sum of its parts.

Specifically, the idea that the hippocampal formation stores representations of stimulus conjunctions is critical to the following important approaches to understanding the hippocampal formation:

- Human amnesia associated with damage to the hippocampal formations has been attributed to the inability to bind together novel stimulus conjunctions (e.g., Marr, 1971; Squire, 1992; Teyler & Discenna, 1986).

- Spatial learning that is dependent on the hippocampal formation has been explained in terms of the ability to acquire a map-like representation of the environment (O'Keefe & Nadel, 1978) or an auto-association process that binds together the stimulus features specific to locations (McNaughton & Morris, 1987; McNaughton & Nadel, 1990).

- Impaired performance in a variety of discrimination learning problems involving ambiguous cues resulting from damage to the hippocampus is said to occur because the subjects cannot use contextual labels (Hirsh, 1974), or acquire configural representations (Schmajuk & DiCarlo, 1992; Sutherland & Rudy, 1989)

- Many computational and/or biologically-based theories of the hippocampal formation emphasize the auto-associative binding properties in area CA3 of the hippocampus (e.g., the Hebb-Marr theory and its descendants; Hebb, 1949; Marr, 1971; McNaughton & Morris, 1987; Rolls, 1989). Related theories emphasize the role of sparseness and conjunctivity in avoiding interference during rapid learning of novel information (e.g. McClelland, McNaughton, & O'Reilly, 1995).

Given that these approaches all incorporate the idea that the hippocampus stores representations of stimulus conjunctions, they would predict that damage to the hippocampal formation should significantly impair performance on problems that require the subject to use such representations. However, several direct tests of the hypothesis that the hippocampus stores stimulus conjunctions indicate that this prediction is *false* (see Rudy & Sutherland, 1995, for a review). Although much of this literature has focused on specifically disproving Sutherland and Rudy's (1989) configural association theory, we argue that these data undermine all of the other theories of the hippocampus that rely on the assumption that the hippocampus stores representations of stimulus conjunctions, creating a major crisis in our understanding of hippocampal function.

In this paper, we will provide a potential resolution to this crisis by considering (a) the capacities of the neocortex to learn in the absence of the hippocampus, and (b) more generally the appropriate dimensions along which the hippocampus and neocortex differ (and the ways in which they are similar).

Many theorists have assumed that the neocortex has relatively limited learning capacities, either implicitly by assuming that cortex without the hippocampus could not learn to represent stimulus conjunctions (e.g., Sutherland & Rudy, 1989), or as an explicit and central part of the theory (e.g., Gluck & Myers, 1993; Schmajuk & DiCarlo, 1992). If, however, the neocortex is recognized as being considerably more powerful (e.g., including the ability to learn conjunctive representations), then the apparent crisis can be resolved. This line of reasoning alone, however, undermines the importance of assigning special significance to the hippocampus in the first place — if the neocortex is so powerful, then what additional benefit does the hippocampus impart?

We think that McClelland, McNaughton, and O'Reilly (1995) have provided an important functional perspective for understanding the division of labor between the hippocampus and neocortex. They argued that the role of the hippocampus is to learn novel information *rapidly*, while the neocortex *slowly* integrates information over many repetitions and thereby represents the underlying (statistical) regularities of the environment. One unified system cannot both rapidly acquire novel information and support representations of environmental regularities because rapid learning produces too much *interference* with previous knowledge. By keeping representations highly *separated* from each other, the hippocampus can learn rapidly while avoiding this interference. However, efficient representation of the structure of the environment requires *distributed, overlapping* representations (which can only be developed through slow learning). This perspective is related to the conjunctive idea in that the separation of representations required for fast learning is thought to take place through

sparse, conjunctive representations (e.g., as explored in a biologically-based computational model of the hippocampus; O'Reilly & McClelland, 1994).

Thus, we argue that stimulus conjunctions can be acquired by two neural systems. One system depends critically on the hippocampus, but also requires the cortical pathways to and from the hippocampus, which we will refer to simply as the hippocampal system. The other system depends primarily on the neocortex and can function without the hippocampal formation. We hypothesize that the operating characteristics of these systems differ in two important ways: (a) learning rate, where the hippocampal system rapidly acquires stimulus conjunctions, whereas the cortical system learns relatively slowly; and (b) bias towards developing conjunctive representations, where the hippocampal system automatically and continuously constructs representations of stimulus conjunctions, whereas the cortical circuit must be driven to construct such representations by the demands of a task, and does not otherwise naturally do so. Thus, the cortex develops conjunctive representations in problem solving situations (e.g., discrimination learning) in which the solution forces the development of conjunctive representations.

This characterization of the differences between learning in the hippocampal system and the cortex makes sense of the recent literature that disproved the Sutherland and Rudy (1989) configural association theory. In all of the cases where no significant deficits from selective hippocampal lesions were observed on conjunctive tasks, these tasks could have driven the remaining cortical system to learn the necessary stimulus conjunctions. Further, many trials of learning were required in both intact and lesioned animals, allowing enough time for slow cortical learning to have acquired the conjunctive representations. In light of this, we conclude that the involvement of the hippocampal system is best revealed by studying tasks where the acquisition of stimulus conjunctions is not forced by task demands (i.e., *incidental* conjunctive learning tasks). This conclusion is supported by several studies which show significant effects of selective hippocampal lesions on incidental conjunctive learning tasks (e.g., Fanselow, 1990; Good & Honey, 1991; Hall & Honey, 1990; Honey, Watt, & Good, 1998; Kim & Fanselow, 1992; Save, Poucet, Foreman, & Buhot, 1992).

In addition to resolving the conflict between the conjunctive theory and behavioral data, our characterization of the differences between the cortex and hippocampus can be used to explain the finding that the hippocampus appears to impart a level of *flexibility* that is not present in hippocampally lesioned animals (e.g., Eichenbaum, 1992). We argue that this flexibility derives from hippocampal *pattern completion* — the ability to complete a hippocampal conjunctive representation from partial cues. Several unique features of the hippocampal formation facilitate pattern completion (McNaughton & Morris, 1987; O'Reilly & McClelland, 1994; Rolls, 1989). In essence, pattern completion enables flexibility by allowing relevant knowledge to be accessed (completed) in novel circumstances. In contrast, the cortex lacks these specialized pattern completion mechanisms, and thus cannot support certain types of flexible behaviors. Nevertheless, we believe that the cortex is capable of other kinds of flexible behaviors that are more consistent with its ability to extract general aspects of environmental structure (e.g., pronouncing novel non-words; Plaut, McClelland, Seidenberg, & Patterson, 1996).

The paper will proceed in several stages. First, we will provide a historical overview of the emergence of the view that the hippocampus contributes to learning and memory by enabling the acquisition of representations of stimulus conjunctions. We will then detail the crisis for this class of theories brought about by the tests of Sutherland and Rudy's (1989) configural association theory. We will then describe in greater detail the proposed solution to this crisis as outlined above, and present a biologically-based computational model of the hippocampal-neocortical system, which instantiates our ideas about the dimensions along which the hippocampus and neocortex differ. This model then will be used to explain the observed patterns of intact and hippocampal lesion data on a wide range of tasks, including nonlinear discrimination tasks, rapid incidental learning tasks, contextual fear conditioning tasks, and hippocampal flexibility tasks.

In summary, we show that a single model of the combined cortical and hippocampal systems can account for a wide range of behavioral data in animals. We focus on the animal data here in part because of its relative simplicity, but the same model has also been used to account for human memory data (O'Reilly, Norman, & McClelland, 1998). Because all of the major aspects of our model can be motivated independently based on computational and biological considerations, it is not merely an *ad hoc* attempt to preserve the conjunctive account in the face of conflicting data, but rather situates this data within a richer overall framework.

Historical Overview

Human Amnesia Studies

It is well known that the story of the hippocampus as a major contributor to human memory begins about 40 years ago with the work of Milner and her colleagues (Milner, 1966; Penfield & Milner, 1958; Scoville & Milner, 1957). Based on extensive neuropsychological examination of a number of patients with unilateral and

bilateral damage to the medial temporal lobes, (most notably, the famous patient H.M.), Milner (1966) concluded that it was damage to the hippocampal formation that was critical to the extensive anterograde and the limited retrograde amnesia that was observed in these patients.

Since Milner's original reports, extensive research has been aimed at characterizing the fundamental deficits common to patients with medial temporal lobe damage and other amnesics. One of the major ideas that has emerged from this research is the view that memory is not a single entity, but rather consists of multiple processes or systems, and that the hippocampal formation is only important for a particular kind of memory (see Squire, 1992 for an interesting review of the history of this development).

The early, more mechanistically oriented accounts of human hippocampal function emphasized the idea that the hippocampus stores stimulus conjunctions (Marr, 1971; Wickelgren, 1979; Teyler & Discenna, 1986). This notion continues to be central as an explanation of how we recall and recognize episodes from our past. For example, this idea was clearly embedded in the memory indexing theory of Teyler and Discenna (1986). They suggested that each experiential event is represented in a unique spatial and temporal array of neocortical modules. By virtue of neocortical-hippocampal information flow, a memory index of the cortical pattern is established in the hippocampus. Subsequently, activation of the memory index by some subset of cues that comprised the original experience will be sufficient to activate the entire array of cortical modules originally activated and provide the basis for recall and recognition.

More recently, Squire (1992) concluded his review with a similar idea of how the hippocampus supports declarative memory. In his words, "In the present account the possibility of later retrieval is provided by the hippocampal system because it has bound together the relevant cortical sites. A partial cue that is later processed through the hippocampus is able to reactivate all of the sites and thereby accomplish retrieval of the whole memory" (p. 224). Note in both of these accounts the hippocampus represents the conjunction of the stimulus features that comprised a particular event or experience and it is the activation of the conjunction that allows memories to be recalled or recognized. These views of hippocampal function correspond well with the notion of *episodic* memory — memory for the specific contents of individual episodes or events (Tulving, 1972, 1983).

In terms of characterizing the contrast between human hippocampal and cortical function, one influential modern view is that the hippocampal formation is important for declarative/explicit memory but not for non-declarative/implicit memory, which can be subserved by the cortex and other brain areas (Squire, 1987, 1992). One important property of declarative memory is that it is accessible to conscious recollection. Among other things, declarative memory enables people to answer questions like: "Have you seen this person before?" "Where did you park your car?" "What did you have for lunch yesterday?" Patients with medial temporal lobe damage cannot answer such questions. Nevertheless, they are able to acquire skills and have their behaviors modified by experience — they possess some other forms of non-declarative/implicit memory (e.g. Graf & Schacter, 1985). The evidence for the distinction between declarative/explicit versus non-declarative/implicit has been reviewed many times (e.g., Squire, 1992, 1987), so we will not do so here. Further, we note that these are descriptive categories of memory and not explanations of how memory works.

Behavioral/Conditioning Studies in Animals

Milner's conclusion that the hippocampus plays an essential role in human amnesia also generated a large volume of animal experimental work. The first wave of studies was summarized in a thorough review by Douglas (1967). From the standpoint of understanding the role of the hippocampal formation in memory, this literature might be considered a major disappointment because, unlike the human literature, the data overwhelmingly demonstrated that rats and primates with extensive damage to the hippocampus and related cortical structures displayed no anterograde or retrograde amnesia.

This epoch, however, contained empirical findings that provide an important link to the development of conjunctive/configural theories. Specifically, Douglas (1967) noted that animals with damage to the hippocampal formation were often impaired in tasks that required the animal to learn a behavior that was incompatible with a previously learned or prepotent response. For example, damage to the hippocampus produced animals that were highly resistant to extinction and slow to learn discrimination reversals (e.g., where the conditioned association is reversed for two stimuli). Based on this pattern of results, Douglas (1967) offered the reasonable hypothesis that the hippocampus was critical to the processes that enable animals to withhold responding — the *response inhibition* view. Douglas, however, realized that a simple view of the concept of response inhibition could not be right. As he put it, "Hippocampally ablated animals do not continue to walk until they bump into a wall..., continue to eat until stuffed..., or continue to groom or scratch for prolonged periods of time once these responses have been initiated. Thus, they are fully able to cease making a response when the initiating stimulus is no longer present" (Douglas, 1967, pp. 434-435).

In refining the concept of inhibition to eliminate this

weakness, Douglas put forth an the idea that contributed to the view that the hippocampal formation stores representations of stimulus conjunctions. Specifically, he suggested that "...the hippocampus might function to inhibit stimulus-response bonds" (Douglas, 1967, pp 435). This refinement is important for two reasons. First, it puts inhibition into a neural system involved in the modulation of associations and thus kept alive the possibility that the hippocampal formation was involved in memory processing of animals. Second, indirectly, he becomes the first theorist to speculate that the hippocampus plays an important role in solving what we will call the *ambiguous cue* problem. This problem emerges because the same stimulus can be associated with fundamentally incompatible outcomes and consequently, a major problem for a mature memory system is to provide a mechanism for solving the problem of associative interference — keeping these different outcome associations separate from each other. Many theorists now assert that the hippocampal formation makes a fundamental contribution to memory by reducing associative interference (e.g., Hirsh, 1974; McClelland et al., 1995; O'Keefe & Nadel, 1978; O'Reilly & McClelland, 1994; Shapiro & Olton, 1994; Sutherland & Rudy, 1989; Rudy & Sutherland, 1994, 1995; Wickelgren, 1979).

Although Douglas alluded to the problem of associative interference and to the potential contribution of the hippocampal formation, he did not provide a mechanism for how this might happen. Here the credit goes to Hirsh (1974), and to Nadel and O'Keefe (1974), O'Keefe and Nadel (1978) as discussed in the next section. Hirsh recognized that the associative interference that accompanies ambiguous cues poses a major problem to animals with damage to the hippocampus. In dealing with this problem, he not only proposed a potential way in which the hippocampal formation could act to reduce associative interference, he also offered what was one of the first examples of a multiple memory view of hippocampal formation function.

Hirsh argued that a learning experience leaves its impact on two different memory systems — the *performance line storage* system, and the *memory* system. These two systems operate quite differently. Generally speaking, experience leaves its effect on the performance line by altering the strength of connection between the neural elements activated by a stimulus and neural elements responsible for the response. Thus, when faced with an ambiguous cue, an organism with only performance line memory must respond solely on the basis of the relative strengths of connection, regardless of whether or not the behavior generated by the strongest connection is appropriate to the task at hand. In contrast, Hirsh's memory system encoded information in a more contextualized fashion in terms of stimulus environments, behaviors, and associated outcomes from the performance line.

It is in Hirsh's view of the operating characteristics of the memory system that a connection between the hippocampal formation and the concept of a conjunctive representation first appears. First, Hirsh argued that damage to the hippocampal formation impairs the operations performed by the memory system. Second, he proposed that the memory system solves the interference problem because it is governed by *contextual* retrieval principles, which insure that context-appropriate associations will be selected and made available to the performance line for action. As Hirsh puts it, "Systems utilizing contextual retrieval do not require deletion of previous learning. The conflicting items of information can be differentiated by the addition of a contextual label indicating that the previously acquired information was formally true" (Hirsh, 1974, p. 426).

For example, suppose the subject is required to learn a biconditional discrimination of the following description. In Context 1 (C1), in the presence of stimulus A the response will be rewarded (+), but in the presence of stimulus B that response will not be rewarded (−). In Context 2, the contingencies will be reversed. We can represent this biconditional discrimination as: $C1 : A+, B-$ and $C2 : A-, B+$. In Hirsh's view, the context is stored in the memory system along with the specific events, and it labels which association is appropriate as a function of the contextual cues ($C1$ and $C2$), thus resolving the otherwise ambiguous meaning of the A and B stimuli. We interpret Hirsh's concept of a contextual label to be equivalent to the use of conjunctive representations that bind together contextual and stimulus features.

The idea that the hippocampal formation contributes to memory by representing stimulus conjunctions emerged unambiguously in a paper by Wickelgren (1979). He argued that the hippocampus is essential to the process of *chunking*. In Wickelgren's words, chunking "...stands for a learning process by which a set of nodes representing constituents (components, attributes, features) of a whole become associated with a new node that there by represents the whole chunk" (Wickelgren, 1979, p. 44). Wickelgren's concept of chunking is clearly equivalent to the concept of conjunctive representations.

The idea that the hippocampal formation stored stimulus configurations (conjunctions) was also embedded in a theory put forth by Mishkin and Petrie (1984) that included many of the same assumptions associated with Hirsh's position. They distinguished between a *habit* and a *memory* system and assumed that the memory system depends on the hippocampal formation and supported the acquisition of stimulus conjunctions.

Spatial Learning in Animals

The idea that the hippocampus stores representations of stimulus conjunctions also emerged in the extremely influential view of the hippocampal formation published by O'Keefe and Nadel (1978) in their now classic (but unfortunately out of print) book, *The Hippocampus as a Cognitive Map*. They also distinguished between two memory systems, a *locale* system and a *taxon* system. Motivated in part by the discovery of place cells in the hippocampus (O'Keefe & Dostrovsky, 1971), they linked the hippocampal formation with the locale system. It supports the acquisition of a map-like representation of the environment, where a map is composed of "a set of place representations connected together according to the rules which represent distances and directions amongst them" (O'Keefe and Nadel, 1978, p. 488). The taxon system is conceptually similar to Hirsh's performance line system as it represents consistent rules, routes, procedures, and stimulus-response habits.

Clearly, the notion that the hippocampus-dependent locale system represents experience as connections between stimulus features (e.g., distance, directions) qualifies it as a stimulus conjunction theory. However, O'Keefe and Nadel (1978) limited the kind of information the locale system could represent exclusively to spatial information in the form of an allocentric spatial map. Their view of the hippocampus has generated an enormous amount of research on both the physiology of the hippocampus and the memory based behavior it supports. Its fundamental behavioral prediction — that damage to the hippocampus will impair performance in spatial learning tasks — has been confirmed many times (c.f., Barnes, 1988).

Not all theorists agree that the deficit in spatial learning associated with damage to the hippocampal formation implies that the subject has lost its ability to acquire a spatial map. The idea that the hippocampus serves as an auto-associator or represents stimulus conjunctions remains at the heart some of these alternative views (McNaughton & Morris, 1987; McNaughton & Nadel, 1990). For example, the patterns of neural firing on an eight-arm radial maze recorded by McNaughton and Barnes (1990) show that neurons in the CA3 fire only in a particular location on a particular arm in only one direction. This apparently map-like encoding could also be explained if the CA3 neurons are activated by particular conjunctions of sensory features that are only present in very specific locations.

Biological and Computational Models

The idea that the hippocampus can represent stimulus conjunctions also emerged independently based on neuroanatomical and computational considerations. As Squire, Shimamura, and Amaral (1989) noted, after the hippocampus was discovered to play a critical role in human amnesia, it became important to have a precise description of hippocampal connections. The reader is referred to Squire et al. (1989) and other detailed summaries of the anatomical and physiological properties of the hippocampus (e.g., Van Hoesen, 1982; Amaral & Witter, 1989; Rolls, 1989; Risold & Swanson, 1996). Here, we will just focus on the the general organizing principles.

Unimodal cortical areas are known to project to polysensory association areas which in turn project to perirhinal cortex and to the parahippocampal gyrus. These structures project to the entorhinal cortex which provides the hippocampus with much of its sensory innervation via the perforant pathway. Thus, the hippocampal formation receives information from virtually all associative areas in the neocortex and "...has available highly elaborated multimodal information which has already been processed extensively along different, and partially interconnected sensory pathways" (Rolls, 1996, p. 607). In addition, to receiving sensory innervation from polysensory associational corticies via the entorhinal cortex, the hippocampus also projects back to these areas via return connections from the entorhinal cortex.

This pattern of connectivity has led a number of theorists to the view that the hippocampus is especially well-suited to represent the pattern of activity or conjunction of specific sensory features of the environment. For example, Rolls (1989) suggests that, "The hippocampus is ideally placed for detecting such conjunctions in that it receives highly processed information from association areas..." (p. 242). McNaughton and Nadel (1990) concluded that, "The activity projected back toward the association cortex by individual neurons can be shown to represent the conjunctions of a broad range of specific sensory features" (p. 25).

A consideration of the computational properties afforded by the hippocampal formation also suggest that it is involved in representing stimulus conjunctions. Many contemporary computational models of the hippocampal formation are influenced by Marr's (1971) theorizing. He sought to infer the computational properties of the hippocampus from its anatomy and physiology. Two of his ideas are especially relevant here. One idea is that the hippocampus provides the substrate for a rapid-storage intermediate/temporary memory system that interacts with a more long-term cortical storage system. The other idea is that it does so as an *auto-associator* — a neural network that can learn to associate the independent elements or components of an stimulus input pattern with each other.

An auto-associator clearly has properties similar to that of a conjunctive representation, as it stores a unitary

representation of a stimulus pattern composed of many separable features. McNaughton and Nadel (1990) note the similarity of Marr's concept of an auto-associator to Hebb's (1949) idea of a cell assembly and refer to such networks as *Hebb-Marr* networks (see also Gluck & Myers, 1997). The idea that the hippocampus serves as an auto-associator and/or represents stimulus conjunctions is a core assumption of a number of contemporary computational models of the hippocampus (e.g., McClelland et al., 1995; McNaughton & Nadel, 1990; O'Reilly & McClelland, 1994; Rolls, 1989).

In summary, this brief review indicates that significant aspects of the behavioral, neuroanatomical, and computational literatures have converged over past 25 years on the idea that the hippocampal formation provides a substrate for representing stimulus conjunctions. This idea emerged early in the history of the field and it is at the core of many contemporary theories of hippocampal function.

Conjunctions in Crisis

Given that the idea that the hippocampus stores stimulus conjunctions has such broad support, it is surprising that there is now a substantial literature that seriously challenges this idea. We will now review the events that led to the current state of affairs.

Of the several theories that have been mentioned, its fair to say that O'Keefe and Nadel's cognitive mapping theory was the most comprehensive and most influential. In spite of its success, however, their idea that the hippocampus only represents topographic spatial relations has never been fully accepted. First, many researchers believe that this idea does not easily account for the basic facts of human amnesia because it is difficult to explain all the facts as a deficit in a spatial map (Hirsh, 1980; Squire, 1992, 1994). Second, in the animal domain, there also exist "exceptions to the rule of space" (Sutherland & Rudy, 1989). Thus, the idea that the hippocampus only stores representations of spatial relationships was thought to be too limited encompass the data.

This state of affairs existed when Sutherland and Rudy (1989) published their *configural association* theory of the hippocampus. Their theory had much in common with the ideas of Hirsh (1974) and Wickelgren (1979). At its core was the assertion the hippocampus was essential to acquisition, storage and retrieval of configural associations. The configural association system combines the representations of the elementary stimulus events to construct unique representations. In other words, it represents stimulus conjunctions.

Sutherland and Rudy's paper is distinguished by two contributions. First, it renewed the challenge to the idea that the hippocampus only represents topographical spatial relations, and showed how a more general idea could in principle explain both the place learning and non-spatial impairments associated with hippocampal damage. Second, and more importantly, Sutherland and Rudy explicitly noted how to provide a strong test of the configural/conjunction theory in nonverbal animals. They argued that there is a set of discrimination problems that are solved by normal animals that require configural associations. The central feature of these problems is that they do not have a linear solution: They cannot be solved by combining the individual associative strengths of component cues that are relevant to the solution.

A prototype example of these *nonlinear* discrimination problems is called *negative patterning*, which is also referred to in the computational modeling literature as the exclusive (X) OR problem (XOR) (Minsky & Papert, 1969; Rumelhart, McClelland, & PDP Research Group, 1986b). Here, the subject is rewarded (+) for responding when either feature A or B is present, but is not rewarded (-) when the compound stimulus AB is present. To solve this $A+$, $B+$, $AB-$ problem, the subject must respond less to AB than to A and B alone. Note that a linear system that can only combine the associative strengths of the elements could not solve this problem because it would always produce more responding to the compound than to the component cues. So, the solution to such problems requires a system that can represent stimulus conjunctions and differentiate conjunctions from their components.

Because nonlinear discrimination problems like negative patterning require a configural/conjunctive representation, and such representations depend on the hippocampus, Sutherland and Rudy made a strong prediction: damage to the hippocampus should impair performance on any discrimination problem that does not have a linear solution. Thus, they provided to directly test the configural/conjunctive theory of hippocampal function.

Given the existing literature, one would have thought that nonlinear tasks would have been extremely sensitive to the effects of damage to the hippocampal formation. Indeed, Rudy and Sutherland (1989) reported that damage to the hippocampus impaired both the acquisition and retention of the negative patterning problem and this result has been replicated by several investigators (e.g., Alvarado & Rudy, 1995b; Sutherland, McDonald, Hill, & Rudy, 1989a). Nevertheless, when Rudy and Sutherland (1995) reviewed the literature generated to provide additional tests of the theory, they were forced to conclude that the strong position they staked out in 1989 could not be maintained. To be sure, there were reports that supported the theory (e.g., Alvarado & Rudy, 1995c; Sutherland, McDonald, Hill, & Rudy, 1989b). More importantly, however, there were clear examples in which damage to the hippocampal formation either

did not prevent animals from solving nonlinear discrimination problems or had no measurable effect (Davidson, McKernan, & Jarrard, 1993; Gallagher & Holland, 1992; Whishaw & Tomie, 1991).

We will only describe two results here and refer the reader to the Rudy and Sutherland (1995) review of this literature. First, Whishaw and Tomie (1991) reported that rats with damage to the hippocampal formation were able to solve a simultaneous biconditional discrimination of the form $AB+$, $CD+$, $AC-$, $BD-$. The stimulus elements were two different diameter strings (A and C) and two odors (B and D). On a trial (e.g., $AB+$ vs. $AC-$) a food pellet was attached to the end of a scented string and the rat was required to pull up the string that contained the food pellet. Second, Gallagher and Holland (1992) reported that rats with damage to the hippocampal formation were not impaired on a discriminated operant problem, $AC+$, $B+$, $AB-$, $C-$ that is very similar to negative patterning ($A+$, $B+$, $AB-$). Their findings were replicated by Alvarado and Rudy (1995b). In each of these cases the damage to the hippocampal formation produced by neurotoxic chemicals was extensive. So, there was little doubt that even without a functional hippocampal formation, rats could solve problems that require a system to represent stimulus conjunctions. Since Rudy and Sutherland's 1995 review, there have been additional reports that the hippocampal formation is not necessary to solve problems that require configural solutions (Bunsey & Eichenbaum, 1996; Cho & Kesner, 1995; McDonald, Murphy, Guarraci, Gortler, White, & Baker, 1997).

The Crisis

Many researchers agree that the literature on nonlinear discrimination problems with hippocampally-lesioned rats provides ample evidence against Rudy and Sutherland's assertion that the hippocampal formation is essential for the acquisition, storage and retrieval of configural/conjunctive representations (Alvarado & Rudy, 1995b; Davidson et al., 1993; Gallagher & Holland, 1992; McDonald et al., 1997; Nadel, 1994; Rudy & Sutherland, 1995; Whishaw & Tomie, 1991). What has not been generally appreciated, however, is that the same data that undermines Sutherland and Rudy's (1989) theory should also be fatal to any theory that assumes that the hippocampus represents stimulus conjunctions.

As we reviewed previously, many different perspectives on hippocampal function share this idea that the hippocampus is uniquely specialized for constructing conjunctive representations. Consequently, the behavioral data that undermines Sutherland and Rudy's qualitative theory appears to be equally problematic for all of these different theories. Without additional argument, it seems unreasonable to assert that the lesion data is fatal to Sutherland and Rudy's position, and yet ignore the implications of the data for all the other theories that make the same assumption as Sutherland and Rudy. Thus, in our view, there is a far-reaching crisis that needs to be resolved to place theorizing about the hippocampus on rational ground.

A Complementary Cortical/Hippocampal Memory System Framework

In this section, we describe a framework that resolves the crisis between conjunctive theories of hippocampal function and the literature suggesting that the hippocampal formation is not always necessary for constructing conjunctive representations. Our approach is to couch the role of the hippocampal formation in establishing conjunctive representations in a broader framework for understanding the division of labor between the cortex and the hippocampus. This framework recognizes that *under well-specified circumstances* the cortex alone can support the acquisition and retention of stimulus conjunctions. It also makes strong predictions about the situations in which hippocampal damage will reliably show behavioral impairments in conjunctive learning. We show that this prediction is consistent with some recent empirical data.

We start by introducing and substantiating some basic principles of cortical function, and then contrast them with principles of hippocampal function. We then show how, in principle, this framework can explain why the hippocampal formation is not necessary for representing the stimulus conjunctions needed to solve nonlinear discrimination tasks. We next elaborate the strong prediction that this framework makes regarding situations where the hippocampus is essential for learning stimulus conjunctions, and describe some of the literature that supports our view. Having provided a general overview of the framework, we then explore an explicit computational implementation of our views and apply this model to a wide range of learning tasks.

Principles of Cortical Function

Various cognitive neuroscience literatures (e.g., electrophysiology, neuropsychology, neuroimaging) suggest that the cortex is responsible for many of the most important and sophisticated aspects of human and animal cognition — object recognition, spatial processing, language, working memory, planning, etc. Furthermore, the cortex is generally regarded as a highly plastic system capable of powerful experience-dependent learning (e.g., when visual inputs are redirected to auditory cortex, neurons there develop characteristic visual receptive field properties; Sur, Garraghty, & Roe, 1988). Thus,

based on these kinds of data, one might conclude that the cortex should be a highly capable system even in the absence of the hippocampal system.

However, several theoretical perspectives and some data suggest that when the medial temporal lobe including the hippocampus is damaged or removed, learning and memory become severely constrained (Squire, 1992), and what is learned has been characterized as highly specific, rote, and inflexible (e.g., Glisky, Schacter, & Tulving, 1986; Squire, 1992; Cohen & Eichenbaum, 1993). In the next two subsections, we summarize two general approaches can be taken towards understanding the learning capacities of the cortex, which we call the *co-dependent cortex* view and the *independent cortex* view. We adopt the later view, and describe it in more detail after first characterizing the co-dependent view.

The Co-dependent Cortex View

The co-dependent cortex view assumes that the cortex can act as a *repository* of knowledge, but that it is not capable of sophisticated learning by itself, and must rely on other brain structures such as the hippocampal formation for acquiring its impressive cognitive functions. One example of this approach can be found in the model of Gluck and Myers (1993). They assume that the hippocampus uses a relatively powerful learning mechanism (error backpropagation), and that the cortex is effectively a slave to this hippocampal mechanism for anything but the most simple forms of learning. The Schmajuk and DiCarlo (1992) model adopts a similar view, as does Rolls (1990) and Wickelgren (1979). They assume that the hippocampus plays an essential role in enabling powerful error-driven learning. In a similar but less computationally explicit account, Cohen and Eichenbaum (1993) attribute the ability to flexibly use acquired information to the hippocampus while maintaining that the cortex is a relatively inflexible subservient system. We also note that the Sutherland and Rudy's (1989) theory assumed that when divorced from the hippocampal formation, cortex was unable to learn difficult nonlinear problems.

The co-dependent cortex view has recently been called into question by several lines of evidence. For example, we have already noted that the animals with damage to the hippocampal formation solve complex nonlinear discrimination learning problems. This suggests that the cortex alone is sufficient for learning these difficult problems. However, perhaps the most dramatic evidence comes from a group of human amnesics who suffered bilateral selective hippocampal damage at relatively young ages (Vargha-Khadem, Gadian, Watkins, Connelly, Van Paesschen, & Mishkin, 1997). Despite having significantly impaired hippocampal function (as was supported by brain scans and evidence of specific episodic memory deficits), these individuals had acquired normal or nearly normal levels of cognitive functioning in language, semantic knowledge, and had normal IQs. Thus, the cortex appears to acquire many of its complex abilities without the assistance of the hippocampus, which is the perspective of the independent cortex view.

The Independent Cortex View and Neural Network Models

The independent cortex view holds that the cortex is a self-sufficient learning system that can develop sophisticated cognitive abilities even in the absence of other systems such as the hippocampus. In addition to the data reviewed previously, this view is consistent with the main body of computational neural network models of human learning, where powerful learning mechanisms have been used to develop models of human language, perception, and many other sophisticated cognitive abilities. These models are typically based on either error-driven backpropagation learning (Rumelhart, Hinton, & Williams, 1986a), or on statistically-based self-organizing learning mechanisms that utilize Hebbian-like mechanisms (e.g., Miller, Keller, & Stryker, 1989). We adopt a model of cortical learning that incorporates both of these forms of learning, providing an overall framework that incorporates a number of important biological, computational, and cognitive principles (O'Reilly & Munakata, in press; O'Reilly, 1998, 1996b).

The critical idea in our cortical model is that representations are shaped through learning by two important aspects of the environment: task demands (via error-driven learning), and the extent to which different events or features co-occur (via Hebbian learning). To enable this information to be useful in novel future situations, only the more general aspects are encoded by the cortex, because the specifics are unlikely to be repeated. Thus, the cortex learns how to generally solve different tasks (e.g., climbing up trees to get fruit), and about which things generally co-occur (e.g., smoke and fire), but not so much about the specific details of a single individual experience. In both error-driven and Hebbian cases, cortical learning abstracts over the details of individual instances and extracts the enduring properties. We will explore some ramifications of this view in the next section. First, however, we will provide a slightly more elaborated view of our model of cortical learning.

Error-driven learning (also called supervised learning) is important for shaping representations according to task demands by learning to minimize the difference between a desired outcome and what the network actually produced (i.e., the error). Backpropagation is a powerful implementation of this idea that uses gradient descent on the error signal to adjust the weights of the

network. Critically, it can use *hidden* layers of units interposed between the input and output units, that enable it to learn nonlinear discrimination problems of the sort discussed above (e.g., the exclusive-or (XOR) problem, that has been studied in the behavioral literature as the negative patterning problem, $A+, B+, AB-$; Minsky & Papert, 1969; Rumelhart et al., 1986a). Thus, if we assume that the cortex is using something like the backpropagation learning mechanism, we would expect that it could indeed learn non-linear discrimination problems without assistance from the hippocampus.

One possible obstacle for thinking that the cortex uses something like backpropagation is that its biological plausibility has been widely questioned. This is primarily because it requires the propagation of error signals in a manner inconsistent with known neurobiological properties (e.g., Crick, 1989; Zipser & Andersen, 1988). Specifically, backpropagation would require in biological terms that an error value be propagated *backwards* from the dendrite of the receiving neuron, across the synapse, into the axon terminal of the sending neuron, down the axon of this neuron, and then integrated and multiplied by some kind of derivative, and then propagated back out its dendrites, and so on. As if this weren't problematic enough, nobody has ever recorded anything that resembles an error signal in terms of the electrical or chemical properties of the neuron.

However, a well-documented property of the cortex, bidirectional activation propagation, can be used to perform essentially the same error-driven learning as backpropagation (O'Reilly, 1996a). The basic idea is that instead of propagating an error signal, which is a difference between two terms, one can propagate the two terms separately as activation signals, and then take their difference locally at each unit. Furthermore, the form of synaptic modification necessary to implement this algorithm is consistent with (though not directly validated by) known properties of biological synaptic modification mechanisms. Also, there are many potential sources for the necessary teaching signals in the form of actual environmental outcomes that can be compared with internal expectations to provide error signals (McClelland, 1994; O'Reilly, 1996a). Thus, it is difficult to continue to object to the use of error-driven learning on the grounds that it is not biologically plausible.

The representation of co-occurance via Hebbian learning mechanisms (Hebb, 1949) is important for forming internal representations (i.e., internal *models*) of the general (statistical) structure of the environment, without respect to particular tasks. We will also refer to this as model learning. Biologically, Hebbian learning requires that the synaptic strength change as a function of the co-activation of the sending and receiving neurons. NMDA-mediated long-term potentiation (LTP) has this Hebbian property (e.g., (Collingridge & Bliss, 1987)). Thus, Hebbian learning is almost universally regarded as being biologically plausible. At a functional level, the co-occurrence of items suggests that there might be a causal relationship between them. Furthermore, co-occurring items can be more efficiently represented together within a common representational structure. Mathematical analyses have shown that Hebbian learning performs something like principal components analysis (Oja, 1982), which extracts the principal dimensions of covariance within the environment. An interesting demonstration of the power of this kind of Hebbian model learning was recently provided in the form of a model that performs principal components analysis on the co-occurrence statistics of words within large texts, yielding surprisingly powerful representations of word meaning (Landauer & Dumais, 1997).

Hebbian model learning and error-driven task learning have complementary objectives, and the combination of both typically performs better than either alone (O'Reilly, 1998; O'Reilly & Munakata, in press). Both appear to be necessary to account for the preserved performance of subjects with damage to the hippocampal formation: Error-driven learning is necessary for learning nonlinear discrimination problems that cortical Hebbian learning typically cannot solve (McClelland & Rumelhart, 1988; O'Reilly & Munakata, in press). However, Hebbian learning can explain phenomena such as preserved repetition priming in amnesics (e.g., Schacter & Graf, 1986), where there are no obvious sources of error or task demands to drive the learning. Although we think that Hebbian learning is an important aspect of cortical learning, much of the remainder of the discussion will focus on the error-driven component, primarily because it enables the cortex to learn conjunctive representations to solve nonlinear discrimination problems.

Limitations of Cortical Learning and the Need for Complementary Systems

Although we believe that the model described above provides a good characterization of the cortex, and that such a cortical system has powerful independent learning abilities, we do not think that it can service all the adaptive functions that the environment requires from organisms. Indeed, the cortical model itself provides some important theoretical leverage for more precisely characterizing the division of labor between the cortex and the hippocampus, by noting where the cortex fails (McClelland et al., 1995).

The failure of standard neural-network models to account for all aspects of human learning was dramatized by McCloskey and Cohen (1989), who noted that a standard error-backpropagation network suffers *catestrophic* levels of interference when applied to a list learning task.

Although many attempts were made to remedy this failure, McClelland et al. (1995) concluded instead that this failure reflects a fundamental tradeoff in learning. On the one hand, the objectives of extracting and representing the *general* properties of the environment must be accomplished. On the other hand, successful adaptation also requires that organisms learn and remember many of the important *specifics* of the world — where you parked your car today, the name of the person you just met, where food or predators were encountered, etc.

These objectives are incompatible, because one representation cannot simultaneously capture both generalities and specifics. Furthermore, the learning mechanisms required to form these different kinds of representations have contradictory properties — acquiring the generalities requires *slow* learning that *integrates* over specific instances, whereas acquiring specifics often requires *fast* learning that keeps the specific instances *separate*. Thus, it makes sense that the brain would employ two complementary learning and memory systems that to optimize these objectives separately. Like McClelland et al. (1995), we believe that the primary role of the cortex is to extract and represent the general features of the environment and the primary role of the hippocampal formation is to represent specifics.

To summarize, when integrating across items to extract generalities, each item contributes a little bit to the representation (i.e., slow learning). When learning about specifics, one needs to keep the items separate from each other, and, because the specifics are more unique and unlikely to recur, one must often learn about them more rapidly. By associating this slow, integrating learning of generalities with the cortex, and the rapid learning of specifics with the hippocampus, we obtain a principled understanding of the division of labor between the two, that is consistent with their respective biological properties, and provides a satisfying interpretation of the learning data.

One key idea that emerges from this division of labor idea is that conjunctive representations are a necessary aspect of rapid, specific learning in the hippocampus, as we will see in more detail in the next section. Thus, we can situate conjunctive representations within the somewhat broader perspective of rapid, specific learning to more precisely characterize the role of the hippocampus and its relationship with the cortex.

Principles of Hippocampal Function

According to the foregoing characterization, the hippocampus should be capable of rapidly learning specific information without suffering undue interference, even when the new information is similar to previously learned information (e.g., where you parked your car today versus yesterday). Thus, the hippocampal system must be able to perform *pattern separation* to keep representations separate and thus avoid interference. We will see in the next section that pattern separation can be achieved by using *sparse, conjunctive* representations, where a relatively few, highly selective units represent a given input pattern.

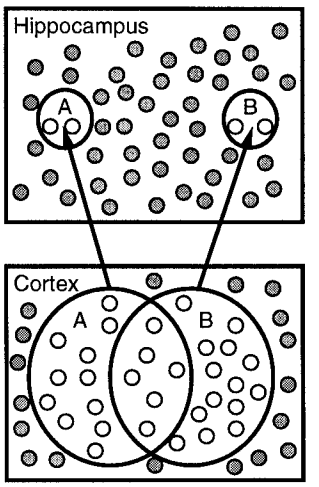

Figure 1: Pattern separation in the hippocampus. Small circles represent units. Circles A and B in the cortex and hippocampus indicate two sets of representations composed of patterns of active units. In the cortex, they are overlapping, and encompass relatively large proportion of active units. In the 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.

However, the hippocampus must also be capable of performing *pattern completion*, where a subset of cues from a previous experience are used to activate or retrieve the memory (stored pattern) of that experience. If only pattern separation were operating, memories could not be retrieved because instead of treating the subset as a retrieval cue, the hippocampus would just store it as a separate pattern. Thus, to actually use the memories stored in the hippocampus, a countervailing pattern completion mechanism is needed. This mechanism is discussed in the subsequent section.

Pattern Separation

Sparse representations (having relatively few active units) lead to conjunctive representations and pattern separation. To understand why, first imagine a situation where the hippocampal representation is generated at random with some fixed probability of a unit getting 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 1). For example, if the probability of getting active for one pattern (i.e., the sparseness) is .25, then the probability of getting active for both pat-

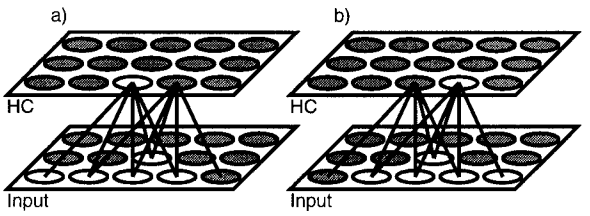

Figure 2: Conjunctive, pattern-separated representations result from sparseness. The extreme case where only one receiving unit (in the upper layer, representing the hippocampus) is allowed to be active is shown here for simplicity. Each receiving unit has roughly the same number of randomly distributed 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.

terns 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 assumes that units are activated at random, ignoring the fact that they are actually driven by weighted connections with the input patterns.

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 2 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. The central idea is that sensitivity to the conjunction of activity in the input produced by a high threshold leads to pattern separation because even if two input patterns share a relatively large number of overlapping inputs, the overall conjunction (configuration) of input activity can be different enough to activate different hippocampal units.

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 2 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.

For pattern separation to work optimally, it is important that different receiving units are maximally activated by different input patterns. This can be achieved by having relatively diffuse, random-looking patterns of partial connectivity with the inputs, which we will see is a property of the hippocampus.

Pattern Completion

Pattern completion is the mechanism that takes a partial input pattern that is a subset of a stored memory, and fills in the missing parts. Thus, when you are asked "where did you park your car today," this input cue is sufficient to trigger the completion of the full encoded memory, enabling you to respond "over by the stadium". Pattern completion is facilitated by particular properties of the hippocampal system, most notably a strong set of lateral connections within a particular layer (CA3) that enable partial activity to spread and fill in the missing pieces.

There is a fundamental tension between pattern separation and pattern completion. Consider the following event: a good friend begins to tell a story about something that happened to them in college. You may or may not have heard this story before, but you have heard several stories about this friend's college days. How does your hippocampus know whether to store this information as a new memory and keep it separate (using pattern separation) from the other memories, or to instead complete this information to an existing memory and reply "you told me this story before." In one case, your hippocampus has to produce a new activity pattern, and in the other it has to produce a old one. If you have perfect memory and the stories are always presented exactly the same way each time, this problem has an obvious solution. However, imperfect memories and noisy inputs (friends) require a judgment call. Thus, there is a tradeoff operating within the hippocampus itself between pattern separation and completion. Optimizing this tradeoff can actually be used to understand several features of the hippocampal biology (O'Reilly & McClelland, 1994), as will be summarized below.

Summary

In summary, these basic ideas about sparse, conjunctive representations for pattern separation, combined with the need for pattern completion, are critical for understanding the functional role of many biological properties of the hippocampus. They provide a basis for un-

derstanding at a mechanistic level why the hippocampus has a strong tendency for creating conjunctive representations, and thus play a central role in our theoretical framework.

It is important to emphasize that although we think the hippocampus is specialized, it still shares many basic properties with the cortex in terms of types of biological learning mechanisms, and the general interconnectivity between and dynamics of excitatory pyramidal neurons and inhibitory interneurons. Thus, we view the hippocampus as a parametrically extreme system that nevertheless utilizes the same basic principles as the cortex — the two systems can be viewed as lying on a continuum instead of as being qualitatively different.

The Junction Between Cortex and Hippocampus

A set of cortical areas provide the interface between the hippocampus and the rest of the cortex, including the parahippocampal, perirhinal, and entorhinal areas. For convenience, we'll refer to these areas collectively as *rhinal* cortex. Until fairly recently, these areas were generally damaged whenever a hippocampal lesion was performed (so it is often called an H+ lesion). With the advent of more selective lesioning techniques, it became possible to lesion only the hippocampus proper, leaving the adjacent cortex intact (a pure H lesion). It turns out that some of the behavioral impairments that had been attributed to the hippocampus proper appear to be more dependent on this adjacent cortex instead (Squire, 1992).

Aggleton and Brown (in press) have recently reviewed a considerable literature that supports the notion that the memory functions of rhinal areas can be distinguished from those of the hippocampus. Specifically, these appear to provide a *familiarity* signal, which reflects recent experience, but, critically, does not contain a full conjunctive, recollectable representation of such experiences. Thus, the ability to perform explicit recall of prior events appears to be unique to the hippocampus proper.

In terms of the nonlinear problem learning literature, this cortical area can be seen as a uniquely privileged higher-order representational area which is one of the few, if not only, places in the cortex where conjunctive representations of multiple elemental stimuli can be formed. Thus, if this area is also damaged along with the hippocampus, especially in the rat, there just isn't any other place where the necessary broad-reaching connectivity is present to enable the formation of novel representations of stimulus conjunctions.

To summarize, we think that the learning and processing properties of the rhinal areas can be understood in terms of basic cortical principles, and not as some kind of grey area between the cortex and hippocampus. However, architecturally, these areas share some of the privileged connectivity of the hippocampus, in that they are interconnected with such a wide range of other cortical areas. Thus, it is not entirely surprising that damage to these areas in addition to hippocampal damage often appears to be just a more severe version of pure hippocampal damage.

Principled Account of Conjunctive Learning

In this section we will describe how our theoretical framework can, in principle, provide an account of performance on tasks that require the learning of conjunctive representations. The principles we use are the major characteristics of the cortical and hippocampal learning systems as just developed. To summarize, the cortex and hippocampus differ on two critical dimensions: (a) learning rate, and (b) relative bias for developing conjunctive representations, but share (c) the ability to learn based on both error-driven and Hebbian mechanisms:

Learning rate. The cortical system learns incrementally and slowly and the hippocampal system 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. The error-driven aspect of learning 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.

We can use these principles to provide a relatively straightforward account of the behavioral data on conjunctive learning. The basic idea is that although the hippocampal system is biased to encode representations of stimulus conjunctions, the cortical system also will store representations of conjunctions if they occur repeatedly

and are part of the structure of environment that must be represented to successfully guide goal directed behavior, or if they predict important events (i.e., via error-driven learning). However, the cortical system will have to learn this information slowly over many trials. Thus, we do not expect the hippocampus to be essential for learning nonlinear discrimination tasks because they are typically acquired over many trials, and under conditions where the contingencies of the environment will engage cortical error-driven learning to acquire the necessary conjunctive representations.

Thus, one important conclusion from our framework is that although nonlinear discrimination problems require that the subject develop representations of stimulus conjunctions, these problems are *not* the most appropriate or most informative ones for revealing the unique contributions of the hippocampal system. So, although we agree with Sutherland and Rudy's (1989) proposal that the hippocampus makes a special contribution to learning and memory because it stores representations of stimulus configurations (conjunctions), we think that the use of nonlinear discrimination problems to test this idea was problematic because these problems do not capture the fundamental way in which the hippocampal system learns stimulus conjunctions.

Our framework suggests that a different class of tasks will be more useful in revealing that the hippocampal formation is fundamentally involved in storing representations of stimulus conjunctions. Our assertion is that the hippocampal system learns stimulus conjunctions rapidly, incidentally and automatically as a function of exposure to environments. In such situations, there is no pressure from the reinforcement contingencies to represent conjunctions. Yet, normal subjects acquire these representations. This rapid, incidental conjunctive learning can be revealed by transfer tasks in which performance is influenced by rearranging the elements of the previous experienced environment. In the next two sections, we will describe several examples of this type that support our view that the hippocampal formation makes an essential contribution to learning these conjunctions.

Rapid, Incidental Conjunctive Learning in Animals

Perhaps the simplest demonstration of rapid, incidental conjunctive learning comes from the study of the role of the hippocampal formation in exploratory behavior. In a well-designed study, Save et al. (1992) repeatedly exposed control rats and rats with damage to the dorsal hippocampus to a set of objects that were arranged on a circular platform in a fixed configuration relative to a large and distinct visual cue. 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. Note that this was not a requirement of the task. All subjects could have habituated simply because they stored representations of the individual objects.

A more recent paper by Honey et al. (1998) makes a similar point. They habituated the rat's orienting response to different sequences of auditory and visual stimuli. On the left side of the apparatus, a tone was followed by the presentation of constantly illuminated light, while the right side had a train of clicks followed by a flashing light. The orienting response to the constant and flashing light habituated in both control rats and rats with excitotoxic hippocampal lesions. However, during a transfer test, in which the auditory and visual combinations were switched (the clicks preceded the constant light and the tone signaled the flashing light) the orienting response to the light was reinstated in the control rats, but not in the rats with damage to the hippocampal formation. Thus, whereas Save et al. (1992) reinstated the habituated response by rearranging the spatial locations of the objects, Honey et al. (1998) reinstated the habituated response simply by altering the stimulus sequence. In both cases, the acquisition of incidental conjunctive representations by the hippocampus, but not the cortex, provides a good account of the data.

There is also evidence from Pavlovian conditioning studies that normal rats learn stimulus conjunctions that are not required by the task. This phenomenon, termed the context specificity effect, is observed in intact rats (Hall & Honey, 1990; Honey, Willis, & Hall, 1990). If rats are conditioned to cue A in Context 1 and cue B in Context 2, they will display more conditioning to cue A in Context 1 than in Context 2 and more conditioning to cue B in Context 2 than in Context 1. Context specificity cannot occur unless the animal stores a conjunctive representation of the cue and the context features, because all the elemental features of the experiment should be equally associated with the unconditioned stimulus (Rudy & Sutherland, 1995). Thus, if responding was just controlled by the associative strengths of the independent elements, there should be no context specificity of conditioning.

Although intact rats display the context specificity effect, Good and Honey (1991) (see also Good & Banner-

man, 1997) reported that rats with damage to the hippocampal formation do not. They respond equally to the cues, regardless of context. Again, it should be emphasized that there is nothing about the original training which demanded that the normal rat learn the conjunctions needed to demonstrate the context specificity effect — they were learned incidentally.

Evidence for the involvement of the hippocampal formation in the incidental learning of stimulus conjunctions has also emerged with the report that rats with damage to the hippocampal formation do not express fear to a context or place in which shock occurred but will express fear to an explicit cue paired with shock (Kim & Fanselow, 1992; Phillips & LeDoux, 1992, 1994 but see Maren, Aharonov, & Fanselow, 1997). This finding has been interpreted by several theorists to mean that the hippocampal formation stores unitary/conjunctive representations of the features that make up the context or place where shock occurred (Fanselow, 1990; Rudy & Sutherland, 1994, 1995). This argument rests on a behavioral analysis of contextual fear conditioning provided by Fanselow (1990). On the surface, there is no obvious reason why damage to the hippocampal formation should interfere with contextual fear conditioning because one might expect that conditioning to one or more of the elements comprising the context could support the conditioned response. Fanselow (1990), however, reached the conclusion that for rats to show normal levels of contextual fear conditioning they must construct a representation the joint occurrence of the elements that compose the context. It is this conjunctive representation that gets associated with shock, and without this representation the animal displays reduced fear of the context.

Fanselow (1990) based his argument on his analysis of what is sometimes called the *immediate-shock effect* (Fanselow, 1986). If animals are given a single strong shock immediately after being placed in the conditioning chambers, they fail to show fear of the conditioning context when tested 24 hours later. However, they do show fear if they are in the conditioning chamber for about 2 minutes before being shocked. Fanselow suggested that animals in the immediate shock group failed to condition because they did not have time to construct the conjunctive representation of the conditioning context before the shock occurred. He provided support for this interpretation by showing that, if the animals were preexposed for 2 minutes to the conditioning context 24 hours prior to conditioning, they would then condition to context even when shock occurred immediately. Presumably, this 2-minute exposure was sufficient to permit the animals to construct the configural representation of the context needed for conditioning when shock occurred immediately after placement in the chamber. Kiernan and Westbrook (1993) replicated Fanselow's results and offered a similar explanation. Thus, we have another example of a task in which normal rats construct representations of stimulus conjunctions simply as a product of exploring and environment.

In summary, the above examples make it clear that there are conditions under which animals automatically acquire representations of stimulus conjunctions just as a natural consequence of being exposed to the environment. Consistent with our theoretical framework, these examples also show that animals with damage to the hippocampal formation do not acquire these representations.

Rapid, Incidental Conjunctive Learning in Humans

Although the human literature provides less definitive evidence, it too is generally consistent with the main prediction of our framework. The primary evidence comes from well known context specificity effects in intact humans, which closely parallels that observed in intact rats. In one dramatic demonstration, Godden and Baddeley (1975) had divers learn a list of 40 unrelated words in one of two environmental contexts: on shore or 20 feet under water. When asked to recall the words in either the same or different context, performance was much better in the same environment than in the different one. This can be interpreted as the effects of the hippocampus automatically forming conjunctive representations that combine together the encoded features of the external environment with the list items.

To identify the hippocampus as being specifically responsible for this incidental contextual encoding in intact humans, data from amnesic patients would be required. A study by Mayes, MacDonald, Donlan, and Pears (1992), showed that global amnesics were not helped by the presence of incidental contextual cues in a recognition memory experiment using word stimuli, whereas normal subjects were. Control and amnesic subjects were matched for performance on recognizing the words without context, so the lack of facilitation in amnesics cannot be attributed to a floor effect. Thus, although the hippocampal localization is not as precise as in the rat studies, it appears that the hippocampus is likely responsible in large part for incidental conjunctive learning in humans.

The generally accepted view that human hippocampal lesions produce specific impairments in *episodic* memory is also generally consistent with our framework. An episodic memory is one that encodes the specific conjunction of environmental and temporal context features together with properties of an event that defines a particular episode (Tulving, 1972). Because such episodes are generally unique, they must be learned rapidly as the episode unfolds. Further, the contextual informa-

tion is typically incidental to any task that might be being performed at the time, yet it appears to be encoded automatically. Perhaps the best available evidence that the hippocampus proper is specifically necessary for this kinds of rapid, incidental, episodic learning comes from the Vargha-Khadem et al. (1997) study of a group of people with early, selective hippocampal damage. They showed that while several forms of learning where repetition and/or non-conjunctive familiarity signals could be used were relatively spared, those forms of learning that required rapid, conjunctive, episodic encodings were impaired.

In summary, the human data appears to be generally consistent with the animal literature in supporting the main prediction of our framework. We will now proceed to demonstrate that this framework can provide an explicit account of both the spared learning of task-dependent conjunctive information by the cortex, and this rapid incidental learning of conjunctions by the hippocampus.

A Computational Neural Network Model

We now describe an explicit computational model that implements the framework outlined in the previous section. The model is based on a computational framework called *Leabra* (O'Reilly & Munakata, in press; O'Reilly, 1998, 1996b), that provides a biologically-based set of activation and learning mechanisms that enable the modeling of both cortical and hippocampal networks within one common framework. The network mechanisms are briefly summarized, followed by a discussion of the architectural properties of the implemented model. Then, we apply intact and hippocampally lesioned versions of the model to a range of learning tasks, and conduct other manipulations to illuminate the basis of its behavior.

Basic Mechanisms

The equations for these mechanisms are all presented in Appendix A, with the main properties summarized here. The basic unit is modeled after the ionic channels present in actual neurons, but the spatial geometry of the neuron has been reduced to a single point. This *point-neuron* formulation maintains close ties to the underlying biology, while remaining nearly as simple as more abstract network formalisms. The modeled units correspond to excitatory pyramidal neurons of both the cortex and hippocampus. The inhibitory interneurons are simulated through the use of a *k-winners-take-all* (kWTA) inhibitory function, which enables a maximum percentage of units (k out of N) to be active at any given time, though fewer than this can be active. This kWTA function approximates set-point negative feedback inhibition from the interneurons, and is implemented by computing a level of inhibitory current that when applied uniformly to all units within a layer allows only k units to be at or above threshold. By setting this k parameter low (e.g., around 5% or less), we obtain the sparse representations of the hippocampal system, and their corresponding conjunctive representations. By setting it higher (e.g., 15-25%), we obtain more integrative, distributed representations characteristic of the cortex.

Learning takes place using two basic mechanisms — a biologically plausible error-driven learning mechanism called *GeneRec* (O'Reilly, 1996a), and a simple Hebbian learning mechanism that has been used in a number of other models (Rumelhart & Zipser, 1986; Nowlan, 1990; Kohonen, 1984). Weight changes are computed by simply adding these two mechanisms together (with a normalized weighting factor).

Overall Architecture and Connectivity

The architecture of the model was designed to capture some very basic and important aspects of the structure of the cortex and hippocampus, while simplifying as much as possible to facilitate analysis of the model's behavior. For most behavioral paradigms, the model learns to associate an input stimulus pattern with an output response pattern, where this response pattern could reflect either the expectation of a reward or punishment, or a specific behavioral response. These input/output associations can be learned both by the cortex (in two different ways) and by the hippocampus, as will become clear.

The overall architecture and connectivity of the model is shown in Figure 3. There are two major components, the cortex and the hippocampus. The cortex includes the basic input/output pathways for carrying out a sensory-motor mapping, including input and response layers that contain simple representations of sensory and motor activity patterns, and three levels of internal representations (elemental, associative, and output). These will be described in greater detail in the next section. The hippocampus interfaces with the cortex via the entorhinal cortex (EC), that captures the information represented in the cortex. The EC then drives the basic anatomical regions of the hippocampal formation, including the dentate gyrus (DG), and the fields of Ammon's Horn, CA3, and CA1. Another input/output area, the subiculum, is not represented here, but is likely to play a similar role to the EC, perhaps with an greater emphasis on subcortical and motor representations. The hippocampal areas form a sparse, conjunctive representation of the entire EC input pattern. Partial input of this pattern can trigger recall of the rest, enabling the hippocampus to take the cortical input pattern and produce an appropriate corresponding output pattern.

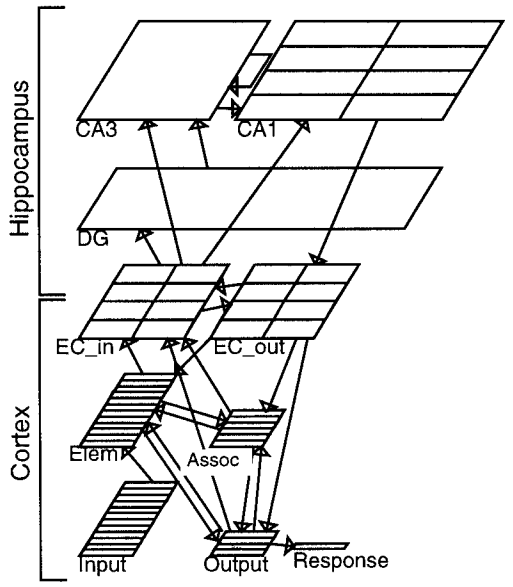

Figure 3: The model, showing both cortical and hippocampal components. The cortex has 12 different input dimensions (sensory pathways), with 4 different values per dimension. These are represented separately in the elemental cortex (Elem). Higher level association cortex (Assoc) can form conjunctive 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.

The Cortical System

All of the representations in the cortical system are organized into groups of 4 units (shown in Figure 3 as the smaller boxes within the cortical layers), with only one out of these 4 units allowed to be active at any given time (yielding a relatively high expected activity level of 25%). This is important for simplifying the interface of the cortex with the hippocampal system as described in the next section. It also simplifies the representational system, and provides a reasonable means of instantiating the tasks that the model will simulate.

The first (elemental) level of internal representation in the cortex is assumed to contain specialized processing pathways that encode information separately along different stimulus dimensions (e.g., different sensory modalities and pathways within modalities such as form, color, or location). Each such pathway is mapped onto a group of 4 units, representing 4 different values along each dimension, and there are a total of 12 such dimensions. Note that values within a dimension are mutually exclusive, but any combination of values across dimensions can be represented. The input simply provides a one-to-one activation of these feature values, but the activations over the elemental layer also reflect the influences from the other layers it is interconnected with.

The association cortex has distributed representations over six active units (one per 4-unit group). Each association unit receives from all of the elemental units, enabling conjunctive representations that combine multiple elemental representations to develop here if required by task demands. This layer may correspond to the rhinal areas in a rat.

Although it typically only represents a binary reward/no-reward value, the output layer also has a population-coded representation over four active units. This distributed output representation is important for providing a sufficiently substantial representation of the output in the hippocampal system, relative to the other cortical areas (which all contribute several active units to the hippocampal input). The output receives full connectivity from the elemental and association cortical areas, in addition to the hippocampal output via the EC. Thus, it can learn a mapping from these areas to a desired output response. Note that because the output receives from all of these areas, each area competes to some extent for influence over the actual output response made.

To more easily decode a binary response from the distributed output layer, the first units in each of the four output groups all project to the first unit in the response, and so on, so that the single unit activated in the response is the one that has received the most "votes" across the four output groups. Thus, the network's behavior is measured as which of the four response units is active.

The cortical areas are all bidirectionally connected, as is consistent with the known biology (e.g., Felleman & Van Essen, 1991). This is important for enabling the biologically plausible GeneRec error-driven learning algorithm to communicate error signals, as described previously. The error-signals in the model come from the difference between an expected reward value over the output layer, and the actual reward value that is received. Thus, the network settles in the expectation phase with the output values updating freely, and then in the outcome phase, the output values are clamped to the actual values. The differences in these two activation states throughout the network are the propagated error signals used in learning.

The Hippocampal System

Our implementation of the hippocampal model is based on what McNaughton has termed the "Hebb-Marr" model (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 mechanisms of pattern separation, learning (synaptic modification), and pattern completion. Further, it relates these mechanisms to underlying anatomical and physiological properties of

the hippocampal formation. Under this model, the two basic computational structures in the hippocampus are the feedforward pathway from the EC to area CA3 (via DG), which is important for pattern separation and pattern completion, and the recurrent connectivity within CA3, which is primarily important for pattern completion. The model relies on the sparse, random, projections in the feedforward pathway from the EC to the DG and CA3, coupled with strong inhibitory interactions within DG and CA3, to form sparse, random, and conjunctive representations. We also emphasize the importance of the CA1 region as providing a means for translating the separated CA3 representation back into the language of the EC, which is necessary to recall information. This can happen if CA1 forms an *invertible* representation of the EC, such that the CA1 pattern can recreate the EC pattern that gave rise to it in the first place (McClelland & Goddard, 1996).

The general scheme for encoding new memories in the hippocampus 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 in these layers. These active units are then bound together in an auto-associator fashion by rapid Hebbian learning within the recurrent CA3 collaterals. Learning in the feedforward pathway also helps to encode the representation. Simultaneously, activation flows from the EC to the CA1, forming a somewhat pattern separated but also invertible representation in CA1. The two different representations of the EC input in CA3 and CA1 are bound together 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 via facilitated connections, which, because it is invertible, is capable of recreating the complete original EC representation. If 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 successful recall.

The rough sizes and activity levels of the hippocampal layers in the rat, and corresponding values for the model, are shown in Table 1. Note that the DG seems to have an unusually sparse level of activity (and is also roughly 4-6 times larger than other layers), but CA3 and CA1 are also less active than the EC input/output layer. The model has very roughly proportionately scaled numbers of units, and the activations are generally higher to obtain sufficient absolute numbers of active units for reasonable distributed representations.

	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, Turlejski, Stanfield, & Cowan, 1987; Boss, Peterson, & Cowan, 1985; Barnes, McNaughton, Mizumori, Leonard, & Lin, 1990)

In a similar manner, the model incorporates rough approximations of the detailed patterns of connectivity within the hippocampal areas (e.g., Squire et al., 1989). Starting with the input, the EC has a columnar structure, and there are topographic projections to and from the different cortical areas (Ikeda, Mori, Oka, & Watanabe, 1989; Suzuki, 1996). This is approximated by the one-to-one connectivity between the cortex and EC. 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 the columnar structure and connectivity of CA1, that is described next.

We noted that for the CA1 to serve as a translator of the pattern-separated CA3 representation back into activation patterns on the EC during pattern completion, it must have invertible representations. At the same time, to minimize interference in the learning of CA3-CA1 mappings, it must also achieve some amount of pattern

separation. Indeed, this pattern separation in CA1 may explain why the hippocampus actually has a CA1, instead of just associating CA3 directly back with the EC input. Thus, the challenge in implementing the CA1 is to achieve both invertibility (which requires a systematic mapping between CA1 and EC) and pattern separation (which requires a non-systematic mapping where similar inputs get mapped to very different representations). This is done in the model by training the CA1-EC mapping to be invertible in pieces (referred to as *columns*), using pattern-separated CA1 representations. Thus, over the entire CA1, the representation can be composed more systematically and invertibly (without doing any additional learning) by using different combinations of representations within the different columns, but within each column, it is conjunctive and pattern separated (McClelland & Goddard, 1996).

The CA1 columns have 32 units each, so that the entire CA1 is composed of 8 such columns. Each column receives input from 3 adjacent EC groups of 4 units (i.e., 12 EC units), which is consistent with the relatively point-to-point connectivity between these areas. The weights for each CA1 column were trained by taking one such column with 9.4% activity level (3 units active) and training it to reproduce any combination of patterns over 3 `EC_in` slots (64 different combinations) in a corresponding set of 3 `EC_out` slots. Thus, each CA1 has a conjunctive, pattern separated representation of the patterns within the 3 EC slots. The cost of this scheme is that more CA1 units are required (32 per column vs. 12 in the EC), which is nonetheless consistent with the relatively greater expansion in humans of the CA1 relative to other hippocampal areas as a function of cortical size (Seress, 1988). A further benefit is that only certain combinations of active CA1 units (within a column) correspond to valid EC patterns, allowing invalid combinations (e.g., due to interference) to be filtered out. We imagine that in the real system, slow learning develops these CA1 invertible mappings in all the columns separately over time.

Finally, we emphasize that although our implementation of the hippocampus is specialized relative to the cortex in terms of patterns of connectivity and levels of activity, the basic processing mechanisms are all fundamentally the same, including the use of inhibitory competition, and a combination of Hebbian and error-driven learning. The inhibitory competition is parameterized differently in the hippocampus and cortex, to achieve the necessary different levels of activity. Also, we have hypothesized that the hippocampus learns incidentally and automatically, so we have set the balance of influence between Hebbian and error-driven learning in the hippocampus to favor Hebbian more strongly. Nevertheless, error-driven learning still plays an important role in the hippocampus, as we'll see when we apply the model to nonlinear discrimination problems.

Application of the Model

We now apply our model to a representative set of findings that are relevant to understanding the role of the hippocampal formation in learning stimulus conjunctions. We first describe simulations of nonlinear discrimination problems, where we find that the model captures the complex patterns of behavior on these tasks exhibited by intact and hippocampally lesioned rats. We then apply the model to problems in which stimulus conjunctions are learned but are not required by the demands of the task. It is in these incidental conjunctive learning tasks where we expect to see the most reliable effects of hippocampal damage. Next, we explore the role of the hippocampus in forming conjunctive representations of context in contextual fear conditioning tasks. In addition to capturing the basic patterns of intact and lesioned behavior, we simulate generalized fear in terms of pattern completion in the hippocampus. Pattern completion also plays a critical role in our final exploration, where we simulate the "flexibility" of hippocampal representations in transitivity tasks.

Nonlinear Discrimination Problems

We begin by reviewing the fundamental property of nonlinear discrimination problems: the individual elements are equally often associated with a rewarded (+) or nonrewarded (−) outcome, so that the problem cannot be solved simply combining the associative strengths of the elements. The negative patterning (XOR) problem discussed earlier is one example, but several other important variations have been employed. These problems can be solved by developing conjunctive representations of the elements, which is what led Sutherland and Rudy (1989) to suggest that these problems should provide a good test of the conjunctive/configural hypothesis.

Although there are some nonlinear discrimination problems where the hippocampus does appear to be necessary (Alvarado & Rudy, 1995c; Dusek & Eichenbaum, 1998; McDonald et al., 1997; Rudy & Sutherland, 1989; Sutherland et al., 1989a), there are others where hippocampal lesions do not impair learning performance (Alvarado & Rudy, 1995b; Gallagher & Holland, 1992; Whishaw & Tomie, 1991). Thus, the simple conjunctive learning story is insufficient to account for the data. An important contribution of our model is to show that it can provide a plausible and principled account of the wide array of results that have emerged from the literature examining the contribution of the hippocampal formation to nonlinear discrimination problems.

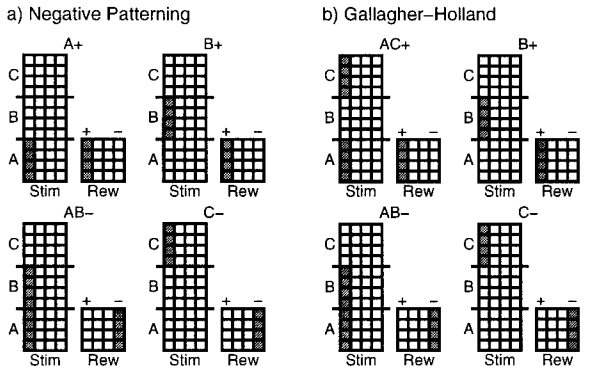

Figure 4: Input/output patterns for the a) negative patterning and b) Gallagher-Holland problems. For each of the four trial types in each problem, the input stimuli (stim) and output reward (rew) are shown. Mutually exclusive values (e.g., + vs - reward) are represented as different values within a dimension, while independent values (e.g., A, B, C) are represented across different dimensions arbitrarily using the first value. The input stimuli in this case are each represented by four dimensions, and the output across six dimensions for reasons described in the text.

Negative Patterning, Gallagher-Holland, and Biconditional problems

We begin by analyzing three problems: (a) the negative patterning problem, $A+$, $B+$, $AB-$, (b) the $AC+$, $B+$, $AB-$, $C-$ discrimination problem introduced by Gallagher and Holland (1992), and (c) a version of the biconditional discrimination, $CA+$, $CB-$, $DA-$, $DB+$. These problems are quite similar and all clearly require the subject to represent stimulus conjunctions. Yet, they are differentially dependent on the hippocampal formation. Thus, this analysis can provide insight into the aspects of these problems that determine whether the hippocampal formation makes a detectable contribution.

First we compare negative patterning (NP) and the Gallagher-Holland problem (GH). There are a number of reports that damage to the hippocampal formation impairs performance on the NP problem (e.g., Alvarado & Rudy, 1995b; Rudy & Sutherland, 1995; McDonald et al., 1997) but has no apparent effect on the GH problem (Gallagher & Holland, 1992; Alvarado & Rudy, 1995b). Indeed, in spite of their similarity, Alvarado and Rudy (1995b) reported that the same animals that were impaired on NP were not impaired on GH.

The NP and GH problems were implemented in the model by presenting the patterns shown in Figure 4. Note that, following Alvarado and Rudy (1995b), we added the $C-$ trial to the NP problem, making it even more similar to GH without changing its logical structure (i.e., the network learns $C-$ very quickly, because it does not conflict with anything at the elemental level). Thus, the only difference between the two problems is the addition of the C stimulus in the $AC+$ trial of the GH problem.

In both cases we compared the performance of the intact model with that of the model with the hippocampal formation component removed (the hippocampal lesion condition). In this case and all subsequent nonlinear discrimination problems, we ran 40 replications with different random initial weights for each condition and the model was trained for 400 epochs (an epoch is one pass through all trial types). The total number of errors was the dependent variable, where an error was defined as a trial inappropriate response. For example, if the model generated a + response on the AB trial, this was an error. Typically, the model made errors until it learned the problem, after which point it performed accurately, so it is possible to interpret this measure as corresponding to the number of trials to criterion. It has the advantages, however, of not requiring the use of a criterion, and of being applicable across different training paradigms (e.g., blocked versus interleaved training, which we explore later).

Figure 5a compares the performance of the intact and lesioned model on the NP problem, and Figure 5b compares intact and lesioned performance in the GH problem. The important finding is that the lesioned model performed worse than the intact model when trained in the NP problem, but the two models were essentially equivalent in the GH problem. Thus, our model is consistent with the findings in the literature — the hippocampal formation is more important for good performance on the negative patterning problem than it is on the Gallagher-Holland problem. We will explain the reasons for the model's behavior in a moment, but first it is useful to consider the biconditional problem.

McDonald et al. (1997) examined the role of the hippocampal formation in several nonlinear discriminations, including the negative patterning problem and a biconditional problem. They found that performance on the negative patterning problem was more impaired by damage to the hippocampal formation than performance on the biconditional problem. In fact, depending on whether one looks at the transformed or non-transformed data, damage to the hippocampus either has no effect or a modest effect on performance on the biconditional discrimination. As discussed previously, Whishaw and Tomie (1991) also found no hippocampal lesion effect in the biconditional.

Our model can also capture this relative sparing of biconditional learning with hippocampal lesions. The stimulus elements in the McDonald et al. experiment were two auditory cues and the presence or absence of a visual cue. Because the auditory cues (A and B) share common features, their similarity was represented by having a 50% overlap in the stimulus patterns that represented their presentation. Similarly, we assumed a 50%

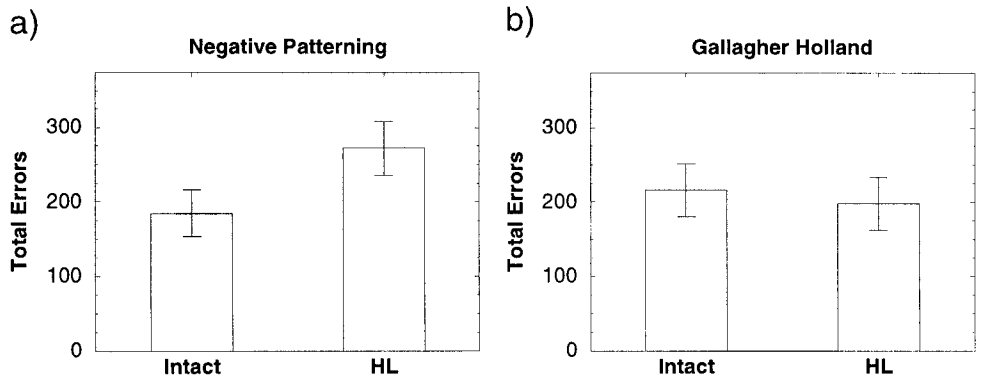

Figure 5: Model results for the a) negative patterning and b) Gallagher-Holland problems. *Intact* is an intact network, and *HL* is a network with the hippocampal component removed ("lesioned"). N=40 different random initializations. Negative patterning is differentially impaired with a hippocampal lesion, due to greater presence of elemental stimuli.

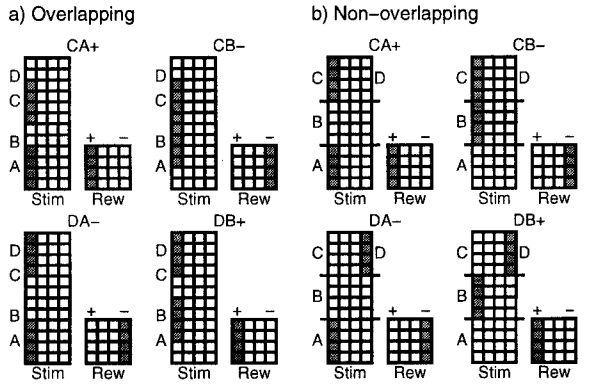

Figure 6: Input/output patterns for the biconditional discrimination problem. a) Shows overlapping case, simulating McDonald et al. (1997), where A and B stimuli overlap 50%, as do C and D. b) Shows non-overlapping case where all stimuli are encoded by different units.

overlap in the input patterns representing the visual cues (*C* and *D*). The Whishaw and Tomie (1991) stimuli were also overlapping (two diameters of string and two odors). To evaluate the importance of stimulus similarity in this problem, we also simulated the case where there was no overlap in the input patterns representing the stimulus elements.

Figure 6 shows the two versions of the patterns we used to implement the biconditional (50% overlapping and completely non-overlapping stimuli). As shown in Figure 7, the intact and lesioned models did not differ when the input patterns representing the stimulus elements overlapped 50%. This finding is thus consistent with the findings of McDonald et al. (1997) and Whishaw and Tomie (1991) indicating that damage to the hippocampus has little if any effect on the biconditional problem. It is interesting to note however that a somewhat different pattern emerged when there was no overlap in the stimulus patterns. Under these conditions, the problem was learned more rapidly overall, and the intact model learned more rapidly than the lesioned one. Thus, the model predicts that the biconditional problem will be learned faster and that it might be possible to detect a hippocampal lesion effect as the stimuli are made more distinctive (e.g., if the stimuli are from different sensory dimensions).

Explanation of the Model's Behavior

Our network clearly replicates the basic findings from the literature — the negative patterning problem is significantly impaired by hippocampal damage, while the Gallagher-Holland problem is virtually unaffected and the biconditional problem is only mildly affected, and in some cases not affected at all. By analyzing the behavior of the model, we should be able to gain insight into the reasons for animal's behavior on these problems. We focus this analysis on two important and related questions: (a) Why does the intact network (and intact animal) require so many trials to solve these problems, even though the hippocampal formation is specialized for rapidly learning conjunctive representations, and (b) How do the differences between these problems interact with the hippocampal and cortical systems such that damage to the hippocampal formation impairs performance on some problems but not others?

Our general answer to the first question is simply that the patterns in these nonlinear problems are *perversely* similar (overlapping), so much so that they stymie the natural pattern separation tendencies of the hippocampus. We noted that the hippocampus has to wrestle with the opposing demands of pattern separation and pattern completion. Thus, if two inputs are sufficiently similar, the hippocampus will have a tendency to perform pattern completion, not pattern separation. So, the notion

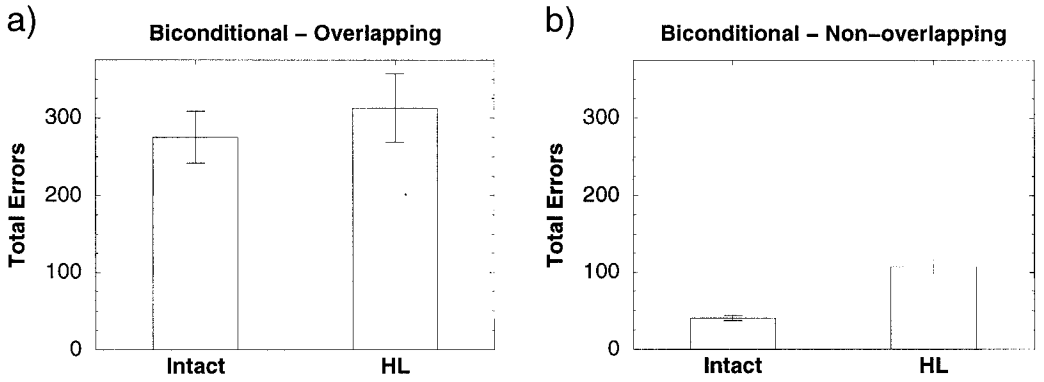

Figure 7: Model results for both versions of the biconditional problem, a) with overlapping stimuli and b) with non-overlapping stimuli. Whether there is an effect of a hippocampal lesion depends on whether the patterns overlap.

that the hippocampus should just automatically separate the patterns in these nonlinear problems is clearly a misconception that fails to appreciate the equally important function of pattern completion.

Given that the hippocampus does not naturally and automatically separate perversely overlapping stimulus inputs, it must rely on the same kinds of error-driven, task-based learning that also shapes the cortical representations. Thus, both the cortical and hippocampal systems must rely on the accumulation of error gradients over multiple presentations to eventually achieve correct performance on these tasks. We think that this is why learning is relatively slow even in the intact network.

Even though the hippocampus and cortex both rely on relatively slow gradient-based learning, the hippocampus still retains an advantage in the form of its more sparse representations and concomitant bias towards pattern separation. This advantage can be seen in both the NP problem and the biconditional problem with non-overlapping stimuli, but not in the GH problem or the standard biconditional with overlapping stimuli. This leads to the second question: What factors lead to the variability in the contribution of the hippocampus to performance across the different nonlinear discrimination problems?

To answer this question, we first identify the essential difference between the three problems (NP, GH, and biconditional). Both Gallagher and Holland (1992) and Rudy and Sutherland (1995) proposed that the essential difference is the extent to which the individual stimulus elements (e.g., A, B, C) appear alone versus in combination with other elements. In the NP problem, both A and B (and C) appear alone, whereas in GH, only B (and C) appear alone. In the biconditional, no elements appear alone.

It is important that stimulus elements appear together because it is impossible to form a conjunction with only

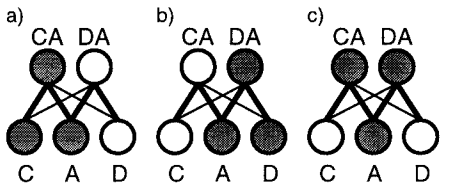

Figure 8: Example of how the presence of multiple stimuli enables the network to easily represent conjunctions in the biconditional problem. If just A is presented, CA and DA are equally activated, but in the presence of C, CA is favored, and in the presence of D, DA is favored.

one stimulus input, and we know that conjunctive representations enable the pattern separation necessary to associate different outcomes with different combinations of stimuli. For example, consider how the network can represent CA separately from DA in the biconditional problem. In Figure 8 we see that there are two representational (hidden) units, that happen to have relatively strong weights to two out of the three input units. Thus, when C and A are presented, the CA unit is differentially excited, and similarly for D and A and the DA unit. However, if we were to just present the A alone, both the CA and DA units would be equally excited. Thus, the presence of multiple input stimuli greatly facilitates the formation of separated conjunctive representations by capitalizing on a kind of *interaction effect* between the two stimuli.

To understand the observed pattern of results, we need to appreciate that the cortex by itself can take advantage of the benefits of multiple inputs for forming conjunctive, pattern separated representations, whereas the extra pattern separation bias of the hippocampus becomes differentially important for solving problems where stimuli appear alone (e.g., NP). In other words, the conjunctive learning capabilities of the cortex are

sufficient for all but the most difficult problems, which strongly reveal the extra contribution of the hippocampus.

It is also possible that biconditional problems that use non-overlapping stimuli will engage the natural hippocampal pattern separation abilities enough to enable the intact model to learn more rapidly than the lesioned model. However, it may be difficult to detect such differences in an experimental context.

To summarize, when compound stimuli (e.g. AB and AC) are present (as in the GH and biconditional problems), the interactions between the elements enable the cortical network to form separated conjunctive representations. However, when a compound stimulus pattern must be separated from its individual elements (as in the NP problem), the extra pattern separation power provided by the hippocampus will be important. In all of these difficult nonlinear problems, many trials are needed because gradient-based error-driven learning is required. However, if the input patterns are sufficiently distinct, it can be possible for the natural hippocampal pattern separation to provide a learning advantage. Thus, the GH and overlapping biconditional problems lie in an intermediate zone of problems with compound, overlapping stimuli, where the hippocampal advantages will be neutralized, and effects of hippocampal lesions will not be reliably detected.

Assessment of Pattern Separation and Blocked versus Interleaved Training

Our analysis suggests that the hippocampus becomes important for solving nonlinear discriminations when the subject is forced by the contingencies of the problem to form separated representations based on elemental inputs (e.g., A and B alone). It also suggests in cases such as the GH problem, where the A element is never experienced alone, that the subject will not have to represent A separately from the AB compound. In contrast, the NP problem forces the subject to represent A separately from AB.

Alvarado and Rudy (1995a) provide evidence relevant to issue of the degree to which the A element is represented separately from the AB compound in these problems. They trained one set of rats to solve the GH problem and another set to solve the NP problem. Then, all rats received several sessions in which they received only $A+$ trials. All rats were then tested on the NP problem. Of particular interest was the effect of the $A+$ training on the rat's response to the $AB-$ compound. If the animals had constructed separated representations of A and AB, then the additional $A+$ trials should have no influence on the rats performance on AB trials — they should be protected from interference. However, if the A representation had not been separated from the AB representation, then $A+$ trials should increase errors on $AB-$ trials. Alvarado and Rudy reported that $A+$ trials dramatically increased errors on AB trials for rats previously trained on the GH problem but had no effect on the errors made by rats trained on the NP problem. This result suggests that rats trained on the NP problem were forced by the task to construct separated representations of A and AB, whereas this was not the case for rats trained on the GH problem.

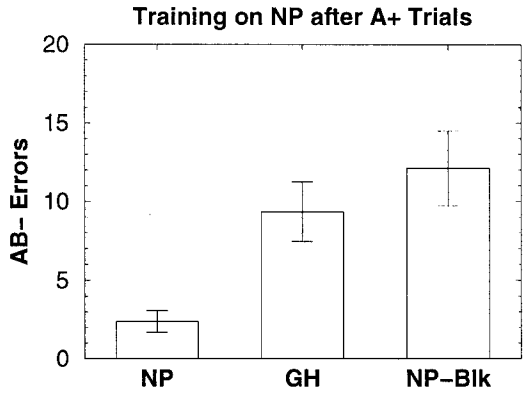

Figure 9: Model results for $AB-$ errors in the NP problem after $A+$ trials. The interference from the $A+$ trials is the least in the interleaved NP problem relative to the other problem types (GH and NP trained in a blocked fashion), indicating that the representation of A is truly separated from that of AB in this case, but not in the others.

We simulated the Alvarado and Rudy (1995a) experiment in our model, and found the same results. As shown in Figure9, additional $A+$ training increased the number of errors on the $AB-$ trials made by rats trained on the GH compared to rats trained on the NP problem.

When trained on the GH problem, both rats and the intact model failed to acquire pattern separated representations of A and AB. These findings support our analysis of why the GH problem is less influenced by damage to the hippocampus than is the NP problem. Clearly, the network (and presumably the animals) are relying on the interaction between A and C in the $AC+$ trials to activate a representation of AC that is separated from the representation of AB. Thus, when A alone is presented, it activates both the AC and AB representations (as shown in Figure 8), resulting in interference for the $AB-$ trials. In contrast, the task demands of the NP problem forces the animal and the model to acquire pattern separated representations of A and AB.

Alvarado and Rudy (1995a) also compared two versions of the negative pattering problem. In one case rats were trained in a standard way: All trial types ($A+$, $B+$ and $AB-$) were pseudo-randomly interspersed in each session. In another case, the rats received blocked presentations of the trial types, with $A+$ trials presented in

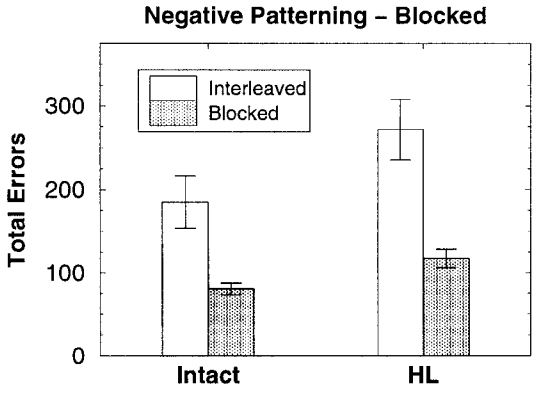

Figure 10: Model results for learning performance in the blocked version of negative patterning for both the intact model and the model with the hippocampal component removed (HL), as compared to the standard interleaved intact and HL data presented earlier.

one block, and $B+$ & $AB-$ trials in another. These rats were then tested on the $A+$ and interleaved NP trials as described previously. Rats in the blocked condition increased their errors (responses) on $AB-$ compounds compared to the standard condition. This result suggests that the blocked NP problem also can be solved without truly separated representations of A and AB.

We also trained the model on the blocked version of the NP problem. Following additional $A+$ training, the model also made more errors on the standard NP problem when it had been trained on the blocked problem than when it had been trained on the standard model (see Figure 9).

We can explain the results of the blocked version of negative patterning by noting that the model reliably made errors at the start of each block, but then rapidly learned (usually within one trial) to produce the appropriate output. Thus, it is clear that the same representation was being used for A and AB, with the mapping between this representation and the response output being rapidly updated for each block (this was confirmed inspecting the representations in the model). This analysis shows that the network must be forced by the task to separate the overlapping representations in these nonlinear problems, and it does not do so if it can minimize errors without separating (e.g., by this rapid re-mapping in the blocked condition). It also supports the idea that the hippocampus in an intact animal is naturally doing pattern completion in these tasks, not pattern separation.

Based on our analysis of how the blocked version of NP is being solved, rats with damage to the hippocampal should not be impaired on the blocked NP problem compared to rats trained to solve the standard NP problem. This is because rats trained on the blocked problem do not have to deal with the difficult task of constructing pattern separated representations of A and AB. The model training results, shown in Figure 10, support this interpretation because there was very little difference between the intact and lesioned model on the blocked version of the problem. This is in contrast with the lesion effect on the standard interleaved version as discussed previously and reproduced in the figure for comparison.

Thus, our model makes a very detailed prediction about the effects of hippocampal lesions on performance in the blocked NP task: Overall error rate should be much lower than in the interleaved, and the differences between the lesioned and intact animals should be concentrated at the beginning of the blocks.

Transverse Patterning

Damage to the hippocampal formation impairs performance on another nonlinear discrimination problem, the transverse patterning (TP) problem (Alvarado & Rudy, 1992, 1995c, 1995b; Dusek & Eichenbaum, 1998). At first glance, this result appears to violate the explanation of why the NP problem is more dependent on the hippocampus than are the GH and biconditional problems, because the TP problem looks like a version of the biconditional problem. However, a more detailed consideration of this problem reveals that it is more similar to the NP problem than the biconditional problem. Thus, the analysis we developed to explain why the hippocampus makes a contribution in the NP problem can also be applied to the TP problem.

An important difference between TP and the other problems we have described is that it requires the subject to make a choice between two stimulus elements. Specifically, the animal has to concurrently solve 3 simultaneous discrimination problems constructed from only 3 elements. Representing the correct choice as + and the incorrect choice as -, we can describe the problems as follows: $A+$ vs. $B-$; $B+$ vs. $C-$, and $C+$ vs. $A-$. Thus, each element is correct or incorrect depending on the other stimulus that is present. The elements could be visual stimuli such as black, white or striped cards (Alvarado & Rudy, 1992, 1995c, 1995b) or odors (Dusek & Eichenbaum, 1998). Typically, the animal is presented with both stimuli, and has to direct a response to one of the elements to indicate their choice.

Because two stimuli are present on each trial, and the correct choice depends on their combination, this task resembles the biconditional. However, the single chosen stimulus is probably in the focus of the animal's attention when the behavioral contingency (reward or no reward) is applied. It is this difference that makes the problem closer to the more difficult negative patterning problem, where stimuli appear individually — revealing the contribution of the hippocampus. Thus, conjunctive representations must be constructed largely from single stim-

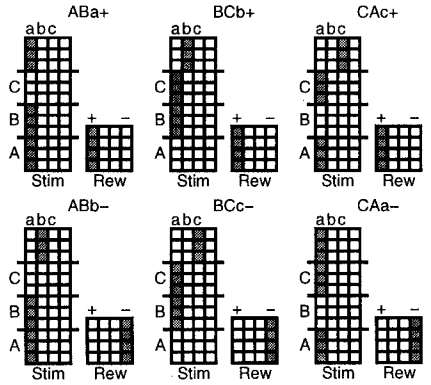

Figure 11: Input/output patterns for the transverse patterning problem. The first set of stimuli (A-C) represent the initial configuration prior to choice, and the second set (a-c) represent which choice was made, with the reward being based on whether the correct choice was made.

uli in the TP problem, and the greater pattern separation bias of hippocampus makes a measurable contribution.

The typical training regime for TP in rats involves three phases. First, they learn the $A+$ vs. $B-$ problem, then the $B+$ vs. $C-$ problem is introduced, and finally the third problem ($C+$ vs. $A-$) is introduced requiring the animal to deal with all three problems in a random mixture of trial types. Note that it is not until the third phase that the problem becomes nonlinear and requires conjunctive processes. Thus, it is interesting to note that rats with damage to hippocampal formation are not impaired until the final phase of the experiment (Alvarado & Rudy, 1995c, 1995b; Dusek & Eichenbaum, 1998).

We implemented transverse patterning in the model in a manner similar to the previous problems. As shown in Figure 11, the network is trained to predict the correct reward associated with making each of the two possible choices in a given trial type (e.g., choosing either A or B in the $A+$ vs. $B-$ trial). We used 3 units in the input space to represent each of the stimuli in the initial configuration (e.g., AB) and 3 units to represent the choice made (e.g., A). Thus, as compared to the biconditional problem, the combination of multiple stimuli is reduced in salience as a result of the space allocated to the choice stimulus. This should make the formation of conjunctive representations more difficult, and therefore increase the dependence on the superior pattern separation bias of the hippocampus.

To test the model, we compared the intact and hippocampally lesioned networks on both the full transverse patterning problem (i.e., all three trial types interleaved), and just the second phase with only two of the three trial types. As is shown in Figure 12, the model captures the pattern of results reported in the literature, with the hippocampal lesion condition impairing performance on the full problem, but not on just the second phase of the problem (which is relatively easy for both the intact and lesioned model, and any differences in performance would not be easily detected in an experimental context). In summary, this problem provides a further confirmation of our previous analysis that having a stimulus appearing alone makes the problem more difficult.

Incidental Conjunctive Learning

We have described how the hippocampus together with the cortex contributes to the solution of nonlinear discrimination problems. We argued earlier, however, that the hippocampal formation makes its most important contribution to memory by automatically and rapidly storing incidental stimulus conjunctions. This function is revealed in experiments on exploratory behavior, incidental learning, and contextual fear conditioning. In this section we apply our model to a representative example of this type of experiment, and in the next section we explore a range of phenomena in contextual fear conditioning.

We noted previously that Good and Honey (1991) provided evidence of hippocampal-formation involvement in incidental learning by studing the context specificity of conditioning. They conditioned rats to cue A in Context 1 an Cue B in Context 2. Normal rats not only condition to the two cues, they also incidently learn where the cues occured because responding to the cues was disrupted if Cue A was tested in Context 2 and Cue B was tested in Context 1. Rats with damage to the hippocampal formation did not display this incidental learning because responding to the cues was independent of the test context.

We applied the intact and hippocampally lesioned model to the context specificity effect to see if it would replicate the Good and Honey (1991) findings. Specifically, we trained the network on two different simple discrimination problems in two different contexts (i.e., $C1$: $A+, B-$; $C2$: $C+, D-$). Although Good and Honey did not explicitly provide nonreinforced cues in each context (i.e., $B-$ and $D-$), their rats were nonreinforced for most of the duration of the training. Thus, we added these nonreinforced cues to neutralize the significance of the contexts. Consequently, although the contexts have no net reward value, the individual stimuli do, and the problem is linearly separable. In effect, one could simply ignore the context, and learn based just on the individual stimuli. However, if the hippocampus is automatically encoding stimulus conjunctions, then a test where the context-stimulus pairs are switched (i.e., $C2$: $A+$, $B-$; $C1$: $C+$, $D-$) should reveal any contribution from such conjunctive representations.

To obtain the relevant data, following training with

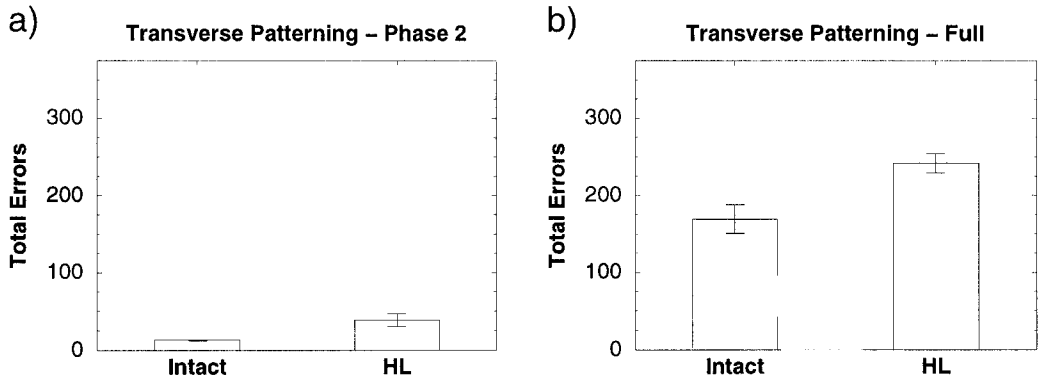

Figure 12: Model results for the transverse patterning problem, for both phase 2 (a), where only two out of the three trial types are used, and the full problem (b), with all three trial types. Only the full problem requires separated conjunctive representations, and it shows an effect of hippocampal lesion (HL) relative to the intact model (Intact). Although the phase 2 effect is statistically significant in the model, the small magnitude of differences involved make it unlikely to find an effect in an experimental context.

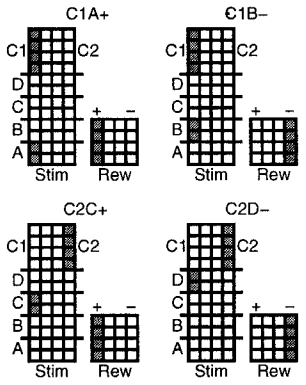

Figure 13: Input/output patterns for the incidental learning context specificity effect. Note cues A and B have equally associative values and that the two contexts C1 and C2 have no net association with reward. If rats respond only to the linear combination of context and cue associative values, then responding should be the same regardless of the context in which the cues are presented.

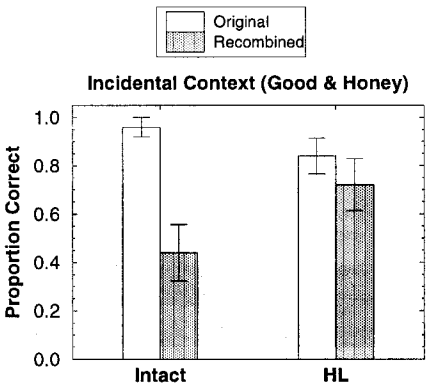

Figure 14: Model results that replicate the Good & Honey, 1991 findings. Shown are the proportion errors during testing with both the original (training) and recombined (switched) contexts. Even though the contexts are completely incidental to the task, the intact model makes errors, whereas the model without the hippocampus (HL) does not.

either the intact or lesioned model, we tested the model under two conditions: (a) the cues were presented in their original contex and (b) the cues were presented in the switched context. The dependent variable was the percentage of correct expectations of the rewards as defined during training. Context specificity then is revealed by the fact that reward outcomes are expected less accurately when the contexts are switched than in the cues are tested in their original training contexts.

The specific patterns we used to train the network are shown in Figure 13. In this and all subsequent simulations, the data are based on 25 replications with random initial weights. Figure 14 shows that the intact model displayed the context specificity effect: Its reward expectations were less accurate when the cues were presented in the switched context than when they were presented in the original context. The model lacking the hippocampus, however, did not display the context specificity effect. It was roughly equally accurate independent of test context.

One interesting parameter that can affect the extent to which the model exhibits the incidental encoding of context is the amount of training time given. For the results shown above, the network was trained to the point where successful performance was achieved. If a longer training period is used, the evidence of conjunctive encoding tends to decrease or go away entirely. This may explain the difficulties that some people have had in obtaining these conjunctive context effects (Hall & Honey,

1990).

Contextual Fear Conditioning

As we noted previously, several researchers have suggested that contextual fear conditioning involves conjunctive representations of the conditioning context (Fanselow, 1990; Fanselow & Rudy, 1998; Maren et al., 1997; Rudy & Sutherland, 1994), and there is evidence the hippocampus makes an important contribution to contextual fear conditioning. In this section we will apply the model to some of the relevant contextual fear conditioning data, showing that the hippocampal system in the model makes an important conjunctive contribution, and that hippocampal pattern completion plays a role in generalized fear conditioning.

The idea that contextual fear conditioning depends on the subject constructing a unitary or conjunctive representation of context first emerged out of Fanselow's analysis of the immediate shock effect. Recall that rats shocked immediately after being placed in the context fail to display fear of that context, whereas rats that experience delayed shock display a substantial fear response. Fanselow (1990) reported that the immediate shock deficit could be ameliorated if the subjects were preexposed to the context prior to the immediate shock session. He argued that context preexposure allowed rats to construct a unitary representation of the context, so that when the rats only briefly encounter a subset of the features on the immediate shock session, the whole pattern is activated and conditioned. We first apply our model to this immediate versus delayed shock effect.

There are three phases to a contextual fear conditioning experiment that must be captured in our model. The first phase is exposure to the context. During exposure, rats explore the environment and presumably are exposed to sequences of stimulus feature conjunctions that, integrated together over time, facilitate the development of a unitary representation of context. The second phase is the delivery of shock. In the third phase the rat is tested by being placed in the conditioning environment, and the percentage of time it spends freezing (exhibiting the fear response) is measured.

In the simulation, we represented the context as four separate stimulus features. We implemented the exposure phase of the experiment by presenting all possible pairwise stimulus feature conjunctions to the network, and allowing it to learn without providing any task inputs (Figure 15). To simulate the kind of temporal integration over individual trials that rats presumably experience, we did not completely reset the activations between trials. Instead, we decayed activations .8 of the way towards 0 from their values in the prior trial. This procedure facilitated the network's ability to form a conjunctive representation of context that integrated over all of the individual features.

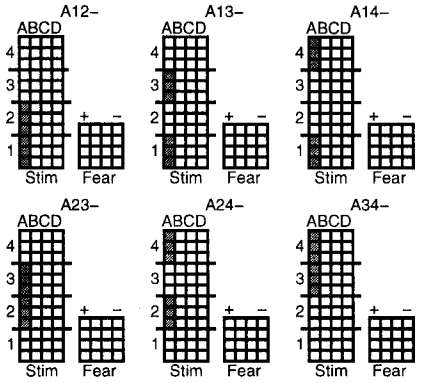

Figure 15: Input/output patterns for the exposure phase of contextual fear conditioning. All possible pairwise combinations of the 4 context features for the A environment are experienced, enabling the hippocampus to encode a conjunctive representation of the fear conditioning context.

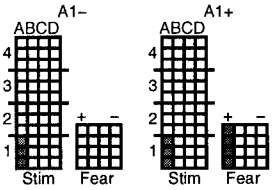

Figure 16: Input/output patterns for the shock phase of contextual fear conditioning. The + output represents a fear response induced by the shock. The input stimulus is assumed to be a single context feature, which is arbitrarily chosen to be the first feature. The fact that the rat views the environment for a brief period prior to being shocked is represented by the initial trial without the fear output activated.

The shock phase was implemented by activating the fear output pattern in the context of a single input feature, representing the fact that the rat receives a relatively narrow view of the environment when shocked (Figure 16). Only a single shock was given. The final phase of fear response measurement was computed as the average fear output activation produced by exposing the network to the sequence of all possible stimulus conjunctions for the conditioning environment (Figure 15). Thus, a strong fear response would be produced if the single shock trial could be associated with a conjunctive representation of context that would be generally activated during testing.

The network was identical to that used previously, with two modifications. The first modification was necessary to ensure that the network did not produce a strong fear response without having first been shocked. This was done by setting the bias weights on the fear output units to -1, a negative bias that must be overcome by learning for these units to become strongly active.

The second modification was necessary to compensate for the fact that the network tends to activate units in the EC layers corresponding to the output layer units even when no external activations to these units are being provided (e.g., in the exposure phase). This has not been an issue previously because the networks were always trained with specific output patterns. However, in this case the spurious activation during exposure causes the network to associate the input stimulus with a non-fear output pattern, which then interferes with the ability of the network to learn the shock-induced fear association during the shock phase. Thus, without supressing these activations, the exposure training has opposing effects — it builds a coherent representation of the context, but it also associates this context representation with a competing output pattern, which interferes with the shock learning[1]. The solution we adopted was to add a negative bias to the appropriate EC units so they will be inactive during exposure.

The first set of simulations demonstrate that the intact model captures the immediate versus delayed shock effect. We compared the level of fear conditioning produced by immediate shock with that produced by exposure to the context for 100 epochs. As shown in Figure 17, the intact model showed a strong level of fear when it was trained for 100 epochs before the shock but almost no fear when it was trained with only a single shock epoch. This exposure facilitation was not evident in the model with the hippocampal component removed, suggesting that the hippocampal system in the model is primarily responsible for the formation of conjunctive context representations.

Preexposure to the context reduces the impaired fear conditioning that results from immediate shock (Fanselow, 1990; Kiernan & Westbrook, 1993). Obviously preexposure to the context would eliminate the immediate shock effect displayed by the intact model because, from the model's standpoint, all that matters is that it be given the opportunity to learn a conjunctive representation of the context prior to the shock — there is no difference between exposure and preexposure in the model.

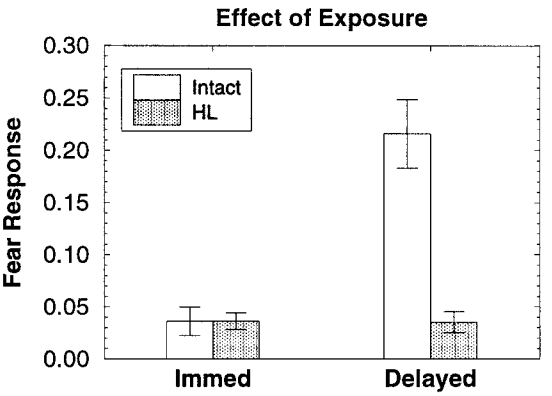

Figure 17: Effects of exposure to the context on level of fear response in the model. Fear response is the activation level for fear output units minus the baseline measure of fear response activation without any conditioning. The immediate shock condition (Immed) is one trial of shock conditioning without any prior training in the environment, showing virtually no conditioning. The delayed shock condition (Delayed) has 100 epochs of training in the environment prior to the shock, and the level of conditioning is strongly elevated in the intact network, but not elevated in the network with a hippocampal lesion (HL).

Is the Representation of Context Conjunctive?

Fanselow and others have assumed that preexposure ameliorates the immediate shock effect because it provides subjects the opportunity to learn a unitary/conjunctive representation of the features that make up the context, but there has been relatively little direct evidence for this assumption. Recently, we provided independent support for this view in a series of fear conditioning experiments with intact rats (Rudy & O'Reilly, submitted). In one experiment, 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 context facilitated contextual fear conditioning, suggesting that conjunctive representations across the context features were necessary. The next simulation shows that the model behaves in a similar manner.

To implement the separate-features condition in our model, we exposed the network to a series of 4 different environments (for 100 epochs each), where each such environment had one of the four conditioning context features (Figure 18). The results of this simulation are shown in Figure 19, which compares the effects of exposure to the elements and exposure to the context with the immediate shock baseline. As in the Rudy and O'Reilly experiment, there was a pronounced facilitation of contextual conditioning when the intact model was exposed to the context as compared to exposure to the features separately. The hippocampally lesioned network showed

[1]This issue of learning a competing output pattern during preexposure affects the extent to which the network exhibits latent inhibition (LI), where context exposure results in subsequently slowed conditioning in that context (Lubow, 1989). One way that LI has been understood, and the way it works in our model, is that a representation of context is being associated with a "no response" representation, which then interferes with the acquisition of the conditioned response (Bouton, 1993). Experience in our own lab has shown that LI is difficult to demonstrate in the contextual fear conditioning paradigm (Rudy & O'Reilly, submitted), and where it has been reported, a considerable amount of preexposure was necessary (Kiernan & Westbrook, 1993). Therefore, the reported results are for complete supression of outputs during exposure, producing no LI effect. However, it is also possible to model a continuum of LI effects by manipulating the activation level of the outputs.

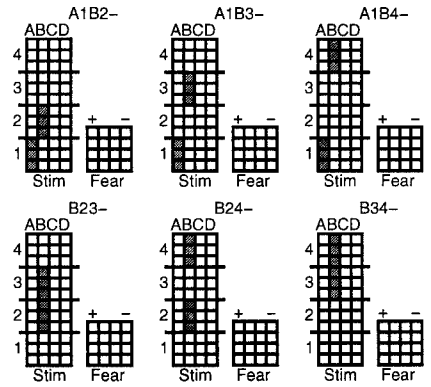

Figure 18: Input/output patterns for exposure to the conditioning context features separately. The first feature of the conditioning context (A1) is mixed in with other features defining a separate environment where this feature was experienced (B2-4). The second conditioning context feature (A2) was similarly experienced in another different environment (C2-4), etc.

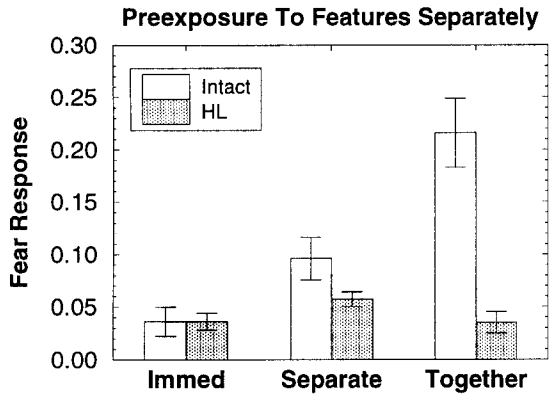

Figure 19: Effects of exposure to the features separately compared to exposure to the entire context on level of fear response in the model (see previous figure for details). The immediate shock condition (Immed) is included for reference. The intact network shows a significant effect of being exposed to the entire context together compared to the features separately, while the hippocampally lesioned network exhibits slightly more responding in the separate condition, possibly because of the greater overall number of training trials in this case.

very little benefit of preexposure to either the context or the features, and if anything responded more in the separate feature exposure condition than in the together condition. This could be due to the greater total number of exposure trials in the separate condition. Thus, as we would expect, the cortex alone does not appear to be sensitive to the stimulus conjunctions in the incidental exposure learning situation.

Pattern Completion and Generalized Fear

An important property of stimulus conjunctions stored in the hippocampus is that they support pattern completion: a subset of an original training pattern can activate the complete pattern. The pattern completion process is central to the contextual fear conditioning phenomena we have just discussed, because it is presumably what enables the testing cues to reactivate the conjunctive context representation and its association with the shock. Recently, we provided novel evidence for the pattern completion processs by studying generalized contextual fear conditioning (Rudy & O'Reilly, submitted). In this section, we show that our model replicates these pattern completion findings.

In Rudy and O'Reilly (submitted), we constructed two contexts, A and B, which shared several features, and a context C that shared none with either A or B. Rats were preexposed to either Context A or Context C, and then conditioned in Context B. Preexposure to Context A should establish an integrated conjunctive representation of that context. Because Context A and B share several features, during the conditioning session, the features common to both A and B should pattern complete to the representation of A, and the A representation will thus become associated with the shock. This means that following conditioning to Context B, rats pre-exposed to Context A will display more generalized fear to A than will rats not preexposed to A (e.g., those preexposed to C). We found that indeed, prexposure to Context A markedly enhanced the rats generalized fear to A. This result strongly supports the idea that rats use a conjunctive representation of the context.

We simulated this experiment in the model by constructing a context A that overlapped with context B by 50% (i.e., shared 2 out of the 4 features), and a context C which overlapped with neither A nor B. Just as in the experiment, the model was then exposed to either A or C (for 100 epochs as before), conditioned in B (with 100 epochs of exposure to B prior to shocking), and then tested in both the A and B environments. The results for the intact and hippocampally lesioned model are shown in Figure 20, which match those of Rudy and O'Reilly (submitted). Preexposure to A and conditioning on B produces an equivalent level of fear when tested on either A or B, but preexposure to C yields less fear in the A test than the B test, because the network did not pattern complete to A when conditioning in B, and thus the A representation did not get associated with shock. However, because there is some level of fear response to A even when preexposed to C, we conclude that the network is also pattern completing somewhat to B in the A testing environment. The lesioned network exhibited a low level of conditioning that did not appear to vary systematically as a function of condition. Thus, we would predict that rats with damage to the hippocampal formation would not reliably exhibit the enhanced generalization effect reported by Rudy and O'Reilly (submitted).

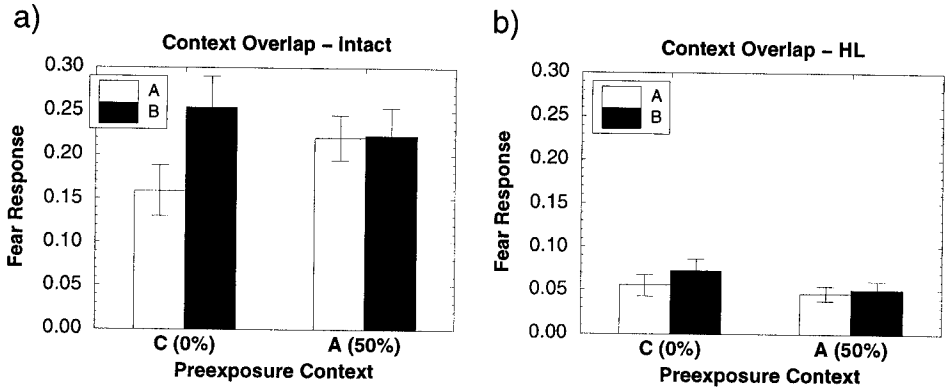

Figure 20: Effects of preexposure to contexts that overlaps with the conditioning context (B) by an amount indicated in the horizontal axis (A has 50% overlap, C has 0% overlap). Testing performed in both A and B contexts. (a) shows the intact network, and (b) shows the hippocampally lesioned network. Pattern completion is indicated in the intact network because the amount of conditioning to A was similar to that shown for B (due to pattern completion based on the 50% overlap). For 0% overlap preexposure (C), A did not get as much facilitation, but still does produce fear, indicating that the effect is a result of pattern completion both at the time of conditioning and at the time of testing. The lesioned network did not show any differentiable effects.

Summary

To summarize, we have been able to account for several of the major properties of contextual fear conditioning using the same basic model that we used on the nonlinear discrimination problems. We see a reliable effect of the hippocampal system in this paradigm because the development of conjunctive representations is not required by the task, and thus the cortical system is not driven to develop such representations. In contrast, the hippocampal system naturally develops these representations, which can be assessed in various ways (e.g., the separate versus conjunctive feature preexposure and pattern overlap conditions as described above).

Transitivity and Flexibility

Several theorists have described memories stored by the hippocampus as being flexible, meaning that (a) such memories can be applied inferentially in novel situations (Eichenbaum, 1992) or (b) that they are available to multiple response systems (Squire, 1992). Some of the best evidence for this comes from studies of *transitivity* in animals (Bunsey & Eichenbaum, 1996; Dusek & Eichenbaum, 1997). In one set of problems, Dusek and Eichenbaum (1997) trained rats to solve a set of concurrent odor discriminations that took the form $A+$ vs. $B-$, $B+$ vs. $C-$, $C+$ vs. $D-$, and $D+$ vs. $E-$. Following training to criterion on these problems, rats were then given probe trials with B vs. D and A vs. E. When confronted with the A vs. E choice both control rats and rats with damage to the hippocampal formation chose A. This is not especially surprising because A was always reinforced and E was never reinforced. The interesting comparison then was how subjects behaved on the transitivity test,

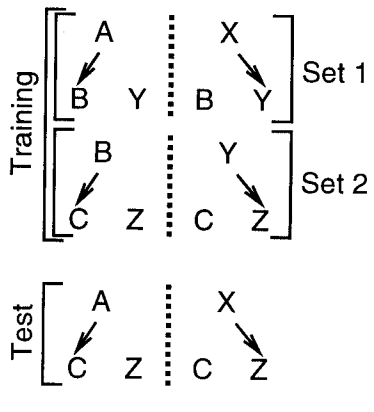

Figure 21: Logic of the Bunsey & Eichenbaum (1996) version of the transitivity test.

the B vs. D probe, because both B and D were equally often reinforced and not reinforced. Control rats consistently chose B, but rats with damage to the hippocampal formation chose randomly.

In the Bunsey and Eichenbaum (1996) version of the transitivity test, rats were trained on two sets of conditional odor discrimination problems (Figure 21). In the first set, they sampled an initial odor (A or X) and then had to choose between two odors (B and Y). When A was the sample the correct choice was B but when X was the sample the correct choice was Y. Then, in the second set, the same rats sampled either odor B or Y (the choice odors of the first set) and had to choose between odors C and Z, where C was correct for sample B and Z was correct for sample Y. After rats had solved these two sets of conditional discriminations, they were given a transitivity test by presenting A and X as samples but

the choice was now between C and Z. Normal rats rat chose C when the sample was A and Z when the sample was X. Rats with damage to the hippocampal system, however, chose randomly.

Eichenbaum and his colleagues argued that the results from both of these experiments support the theory that the flexible nature of hippocampally-mediated memories enables the rats to perform a kind of logical inference. In the Dusek and Eichenbaum (1997) version, they argued that the rats apply a transitivity operation to the B vs. D case, and infer that because $B > C$, and $C > D$, that it must be that $B > D$. Specifically, they propose that their rats had stored the problems as an orderly hierarchy that includes all 5 elements of the 4 problems ($A > B > C > D > E$) that can be used flexibility in the service of supporting logical inferences. Similar arguments were made regarding the Bunsey and Eichenbaum (1996) version.

Our analysis of the two tasks used to demonstrate transitivity suggests that both results are a product of the pattern completion properties of the hippocampus, not the use of logical reasoning. Furthermore, our account shows that the detailed means for achieving transitivity in these two tasks are somewhat different, and that both depend critically on the specific training procedures used. Both tasks depend on hippocampal pattern completion to activate a representation developed during the training procedure that produces the correct transitivity response. Because the transitivity test probes (B vs. D in Dusek & Eichenbaum and AX, CZ in Bunsey & Eichenbaum) overlap with multiple training patterns, producing the correct transitivity response requires that a specific hippocampal representation be favored in this pattern completion process over other possible such representations that also overlap with the test probes. We show below that the two tasks differ in the way that this specific hippocampal representation is favored as a function of the training parameters.

The $A > B > C > D > E$ Transitivity Problem

The key to understanding how the rats solve the Dusek and Eichenbaum (1997) transitivity test is in the training procedure. They trained the rats in ordered trial blocks, starting with 10 trials on the $A+$ vs. $B-$ problem always followed by 10 trials on the $B+$ vs. $C-$ problem, always followed by 10 trials on the $C+$ vs. $D-$ problem and so on. Over the course of training, the number of trials per block was reduced gradually to the point of single trials of each type, and then randomly interleaved trials were run at the very end. This training likely caused nearby trial types in the in the $A > B > C > D > E$ sequence to have overlapping hippocampal representations, because each problem overlaps 50% with the next one, so it is likely that some hippocampal units exhibited pattern completion and were activated for the two adjacent trial types.

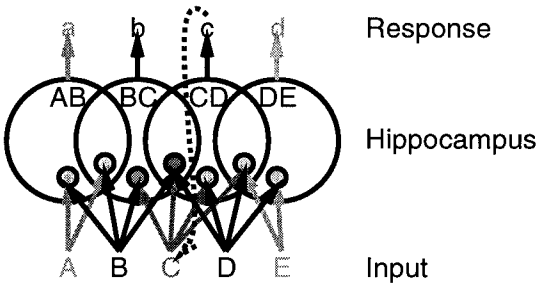

Figure 22: Illustration of how overlapping hippocampal representations can lead to correct transitivity response for the B vs. D probe. The large circles each represent the collection of hippocampal units encoding a given comparison, as labeled (e.g., AB is $A+$ vs. $B-$). The overlap in representations is shown as overlap in these circles. Representative units from each region are shown as small filled circles, with the activation of each unit indicated by the darkness of the circle. The B vs. D probe preferentially activates the overlapping region between the BC and CD representations, because units in this region receive from both B and D inputs, while units in all other regions only receive from one input. The pattern completion property of the hippocampus will tend to complete to either the BC or CD representation, and activate the corresponding response output (B or C, respectively). The C response, not being a valid option for the B vs. D probe, will instead activate the input representation of C, which will then bias the network in favor of completing to BC instead of CD, and thus making the correct response (B).

As Figure 22 shows, the overlapping hippocampal representations can then activate the correct B response for the B vs. D probe via pattern completion. Specifically, if the hippocampal representations for $B+$ vs. $C-$ (BC) and $C+$ vs. $D-$ (CD) overlap, then the overlapping portion of these representations will be activated by both B and D in the B vs. D probe. Due to pattern completion, one of the two hippocampal representations will be activated (BC or CD), and produce the corresponding response (B or C, respectively). However, because C is not available as a choice option on the B vs. D probe, the rat is unlikely to make use of the CD representation directly. Instead, it is likely that the C response will trigger the representation of C as an input, which would then favor the activation of the BC hippocampal representation, producing the correct B response to the B vs. D probe.

To evaluate this account in our model, we first pretrained the network to associate responses with input stimuli (e.g., so that the C response will preferentially activate the C input representation via the pre-existing bidirectional connectivity between them), which we assume the rat would naturally do. Then, we trained the

always before DRINKing it is first necessary to OPEN the container. The model treats necessary and plausible links in different ways. Instruments are connected to events (actions) via the instrument (INSTR) link type. For example, to express 'pounding something with a hammer', (POUND, ?AGT, ?OBJ) is connected to (USE, ?AGT, HAMMER) by the instrument (INSTR) link, where ?AGT and ?OBJ denotes an unspecified agent and object. (see also Figure 4).

Connection strengths for each link are free parameters. They could be estimated based on some empirical data. In the present study, however, the main objective for the modeling work is to provide qualitative simulations for the experimental results rather than quantitative predictions, and thus a priori values are used and then adjusted if necessary to produce desirable outcomes.

Simulation Walkthrough

We will walk through a sample simulation to illustrate the retrieval and construction-integration processes. The following text is used for illustration:

3.3. John took the hammer out of the garage. He pounded the boards together in the afternoon.

In this study, parsing processes are excluded from consideration. Analysis of the text begins with propositional representations which are assumed to be constructed by the parsing component of the system (Kintsch, 1974). The text is analyzed into the following propositions:

P1 (TAKE , JOHN, HAMMER)

P2 (LOC, PROP P1, GARAGE)

P3 (POUND, JOHN, BOARD)

P4 (TIME, P3, AFTERNOON)

The propositions, P1 and P2, come from the first sentence, and P3 and P4 from the second sentence. It is assumed that the basic unit of processing is a sentence. Hence, P1 and P2 are processed together first while P3 and P4 are processed together in a later processing period.

STEP 1:

The elaboration process was performed after processing all the propositions from the text since it is legitimate to assume that the comprehender does not engage in extensive elaboration of a text until all sentences are read especially when the length of the text is very short and in an experimental setting where the comprehender usually does not have a motivation for elaboration. The initial propositional network is constructed with the propositions P1 and P2, which are connected to each other with a bi-directional link with the strength of 1.0. The propositions P3 and P4 are processed in the same manner as P1 and P2 are processed. In the network shown in Figure 5, the black nodes and the connections among them are constructed here.

Now the elaboration process begins; the text propositions retrieve pieces of knowledge related to them from LTM. In order for a proposition to serve as a retrieval cue for elaboration, it needs to be a node newly added to the network and must have the activation value equal to or greater than the threshold. In this example, the threshold is set to 0.4. The four text propositions in the current network qualify because they are new propositions and their activation values are 1.0. These nodes retrieve the long-term memorypropositions P5 (USE, JOHN, HAMMER), P7 (TOOL, HAMMER, POUND)[2], P8 (FIND, JOHN, HAMMER), and P9 (HAVE, JOHN, HAMMER), which are connected to their retrieval cue propositions

[2]Not only propositions but also arguments within the propositions can serve as retrieval cues. In this case, the object HAMMER in P1 retrieves (TOOL, HAMMER, POUND), which represents that HAMMER is a TOOL for POUNDing.

with the link types specified in the LTM, and assigned the activation value of 0.0 and the self-strength of 1.0. For example, P9 (HAVE, JOHN, HAMMER) is connected to P1 (TAKE (AGT JOHN) (OBJ HAMMER)) with the link type of CONSEQ, and it denotes that John took a hammer and as a result he had it. Note that P3 (POUND, JOHN, BOARD) has retrieved P5 (USE, JOHN, HAMMER) and P6 (NOT (USE, JOHN, HAMMER))[3] and they are connected to each other by an inhibitory link. The negation proposition P6 thus competes with P5. This means that given (POUND, JOHN, BOARD), it is probable that he used a hammer, but he may not have. However, the connection strength between (POUND, JOHN, BOARD) and (USE, JOHN, HAMMER) is twice as great as that between (POUND, JOHN, BOARD) and (NOT (USE, JOHN, HAMMER)). This indicates that there is a bias toward the affirmative proposition; namely, it is more likely that one assumes that John used a hammer than one assumes otherwise. Then the network goes through the integration phase. In the construction phase of the next step, those long-term memorypropositions that have been added to the network in Step 1 are potential retrieval cues. In fact, propositions P7, P8, and P9 have gotten the activation values greater than the threshold of 0.4 as a result of integration, and thus serve as retrieval cues (marked by a thick circle in Figure 5).

STEP 2:

The sytem further elaborates the representation with P7 (TOOL, HAMMER, POUND), P8 (FIND, JOHN, HAMMER), and P9 (HAVE, HAMMER) as retrieval cues in Step 2. P7 (TOOL, HAMMER, POUND) connects itself to P5 (USE, JOHN, HAMMER) when it retrieves (USE, JOHN, HAMMER). Note that connections between P1 and P8, and P1 and P9 are now bi-directional. This is because P8 (FIND, JOHN, HAMMER) and P9 (HAVE, HAMMER) retrieve (TAKE, JOHN, HAMMER)

[3]The knowledge base does not contain negation propositions. A negation proposition such as this one is generated during the construction phase.

and make connections to it. As a result of the construction process in Step 2, the network as shown in Figure 5 is built in the working memory. Again, the network is submitted to the integration phase. When the integration process is done, the system checks the activation values of those newly added propositions (i.e., P10 through P15), and finds that none of them has gained an activation value equal to or greater than the threshold. Thus, no more knowledge is retrieved from LTM, and the processing is completed.

Insert Figure 5 about here.

Simulation Results

Several simulation runs were conducted with different texts. The main goal of these simulations is to provide an adequate account for the experimental data of the present study. For this purpose, simulations were intended to be qualitative in that hey showed the major trends in the experimental results, and extensive parameter estimation was not attempted.[4] As in the experiments, texts were constructed by combining each of the four context sentences and the action sentence as shown below:

3.4a. John broke the hammer in the garage. (Contradictory)

3.4b. John looked for the hammer in the garage. (Weak)

3.4c John found the hammer in the garage. (Moderate)

3.4d. John took the hammer out of the garage. (Strong)

3.5. He pounded the board in the afternoon. (Action)

---

[4]Of course, there are drawbacks of qualitative simulations. For example, there is no statistical evaluation, and thus it leaves room for disagreement about validity of a model (Kintsch, 1992).

Simulations of Experiments 1 and 2

Experiments 1 and 2 showed that instrument inference was drawn only when the context strongly supports the inference. We assume that in these experiments the participants did not have any specific goals and hence were not motivated to adopt an elaborative reading strategy. In the simulation, this assumption is realized by setting a high activation threshold (0.4). The simulation run with the strongly related context has been shown in the above walkthrough (see Figure 5). The contradictory text and the moderately related text were also submitted to simulation in the same manner as the walkthrough.

Figure 6 shows the time course of activation of the instrument proposition for the strong context (i.e., the TAKE-HAMMER text) and moderate context (i.e., the FIND-HAMMER text). For the first ten processing cycles (Step 1) the activation patterns are identical for both the strong and moderate contexts. This is because in Step 1 of processing, for both contexts, the inference proposition is retrieved by the action proposition (POUND, JOHN, BOARD) and receives activation from it. However, in Step 2, for the strong context, the propositions (HAVE, JOHN, HAMMER) and (TOOL, HAMMER, POUND) make connections to the USE-HAMMER proposition and send activation to it, and as a result a higher activation is achieved (see Figure 5).

Insert Figure 6 about here.

On the other hand, for the moderate context, while the TOOL proposition sends some activation to the USE-HAMMER proposition, (HAVE, JOHN, HAMMER) does not make a connection to the USE-HAMMER proposition because it does not reach the activation level high enough to be a retrieval cue (see Figure 7).

As a result, the activation steadily decreases. When the networks were settled after 24 processing cycles, the activation of the inference for the strong context was higher than that for the moderate context. In Experiment 2, the results showed the mean reaction time for the strong context was faster than that for the moderate context. Thus, the model has successfully simulated the results.

Insert Figure 7 about here.

The simulation run with the contradictory context exhibited a different pattern of behaviors. The activation of negation proposition (NOT (USE, JOHN, HAMMER)) shot up quite rapidly and kept a high value throughout the processing, whereas (USE, JOHN, HAMMER) received some activation initially but lost it quickly. This is because the NOT-USE-HAMMER proposition is connected to the BREAK-HAMMER proposition and receives activation and wins the competition against the USE-HAMMER proposition, which receives only much smaller amount of activation from the POUND-BOARD proposition due to the smaller link strength for the INSTRUMENT link. Thus, the simulation predicted that for the contradictory condition, the instrument inference would not be drawn. Indeed, Experiment 1 found the instrument priming for the strongly related context but not for the contradictory context.

Simulations of Experiment 3

In Experiment 3, the participants were trained to attend to the use of the instrument by answering comprehension questions about instruments, and instrument priming was observed for both the moderate and strong contexts. This result showed that when the comprehender was engaged in a deeper processing to

make more elaboration, instruments were inferred as long as the context is compatible with such an inference.

To simulate these results, the activation threshold was lowered to 0.1 in the subsequent set of simulations. This made the system pay attention to and make use of those propositional nodes with lower activation, which did not function as retrieval cues in the earlier simulations. The same sets of texts were used for these new simulations. Figure 8 shows the comparison of the final activation of the inference between the high threshold and the low threshold simulation runs for each text. For the strong context, the inference achieved high activation for both thresholds. For the weaker contexts, where the activation of the inference was low for the high threshold, higher level of activation resulted when the threshold was lowered. For the contradictory context, the change in the threshold did not have any effect; namely, the inference node was not activated at all in either case. These outcomes are consistent with the experimental results.

Insert Figure 8 about here.

A closer look at the model's comprehension process for the FIND-HAMMER text reveals more details about how these results came about. The activation patterns for the inference proposition in both simulation runs look identical during the first two steps of the processing. However, in Step 3, the activation increased because the HAVE-HAMMER proposition, which had been added to the network in Step 2, made a connection to the USE-HAMMER proposition and sent activation to it (compare Figures 7 and 9). This indicates that the integration of the proposition (HAVE, JOHN, HAMMER) into the representation is crucial in activating the inference proposition because it functions as the major source of facilitation for the

inference. In other words, as discussed in the experiment section, the presence of the HAVE-HAMMER proposition serves as a precondition for the inference to be drawn.

Insert Figure 9 about here.

---

4. SUMMARY AND CONCLUSIONS

In the present study, we hypothesized that the structure of a situation model influences elaborative instrument inferences during comprehension. The experimental results give support to the hypothesis. The results showed that under normal reading without any specific reading goals, the degree of elaboration involved in construction of a situation model is not extensive. However, if the comprehender is motivated to elaborate and retrieve more knowledge during the process, the inference is computed even in the context that weakly supports the inference. The CI model successfully displayed the behaviors that are qualitatively in agreement with the experimental results reported in the previous section. The model provides an explicit account for the inference computation processes and makes predictions about such processes.

In conclusion, On-line inference processes, particularly elaborative inferences, are influenced by the trade-off between the demand for constructing a rich situation model and the limited capacity of the human cognitive system. On the one hand, the comprehender tries to activate and integrate as much world knowledge as possible to construct a situation model as much elaborated by world knowledge as possible. On the other hand, due to the limited capacity, the comprehender can allow only a limited amount of resources for the on-line inference processes. As a

References

Aggleton, J. P., & Brown, M. W. (in press). Episodic memory, amnesia, and the hippocampal-anterior thalamic axis. *Behavioral and Brain Sciences.*

Alvarado, M., & Rudy, J. W. (1992). Some properties of configural learning: An investigation of the transverse patterning problem. *Journal of Experimental Psychology: Animal Behavior Processes, 18,* 145–153.

Alvarado, M. C., & Rudy, J. W. (1995a). A comparison of configural discrimination problems: Implications for understanding the role of the hippocampal formation in learning and memory. *Psychobiology, 23,* 178–184.

Alvarado, M. C., & Rudy, J. W. (1995b). 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.

Alvarado, M. C., & Rudy, J. W. (1995c). Rats with damage to the hippocampal-formation are impaired on the transverse-patterning problem but not on elemental discriminations. *Behavioral Neuroscience, 109,* 204–211.

Amaral, D. G., & Witter, M. P. (1989). The three-dimensional organization of the hippocampal formation: A review of anatomical data. *Neuroscience, 31,* 571–591.

Barnes, C. A. (1988). Spatial learning and memory processes: The search for their neurobiological mechanisms in the rat. *Trends in Neurosciences, 11,* 163–169.

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.

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.

Bouton, M. E. (1993). Context, time, and memory retrieval in the interference paradigms of pavlovian learning. *Psychological Bulletin, 114,* 80–99.

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

Cho, Y. H., & Kesner, R. P. (1995). Relational object association learning in rats with hippocampal lesions. *Behavioral Brain Research, 67,* 91–98.

Cohen, J. D., & O'Reilly, R. C. (1996). A preliminary theory of the interactions between prefrontal cortex and hippocampus that contribute to planning and prospective memory. In M. Brandimonte, G. O. Einstein, & M. A. McDaniel (Eds.), *Prospective memory: Theory and applications.* Mahwah, New Jersey: Lawrence Earlbaum Associates.

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.

Davidson, T. L., McKernan, M. G., & Jarrard, L. E. (1993). Hippocampal lesions do not impair negative patterning: A challenge to configural association theory. *Behavioral Neuroscience, 108,* 227–234.

Douglas, R. J. (1967). The hippocampus and behavior. *Psychological Bulletin, 67,* 416–442.

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

Dusek, J. A., & Eichenbaum, H. (1998). the hippocampus and transverse patterning guided by olfactory cues. *Behavioral Neuroscience, 112,* 762.

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

Fanselow, M. S. (1986). Associative vs. topographical accounts of the immediate shock-freezing deficit in rats: Implications for the response selection rules governing species-specific defensive reactions. *Learning and Motivation, 17,* 16–39.

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

Fanselow, M. S., & Rudy, J. W. (1998). Convergence of experimental and developmental approaches to animal learning and memory processes. In T. Carew, R. Menzel, & C. Shatz (Eds.), *Mechanistic relationships between development and learning: Beyond metaphor. dahlem workshop report.* (pp. 243–304). Clichester: John Wiley & Sons.

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex, 1*, 1–47.

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.

Gluck, M. A., & Myers, C. E. (1997). Psychobiological models of hippocampal function in learning and memory. *Annual Review of Psychology, 48*, 481–514.

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.

Good, M., & Honey, R. C. (1991). Conditioning and contextual retrieval in hippocampal rats. *Behavioral Neuroscience, 105*, 499–509.

Graf, P., & Schacter, D. L. (1985). Implicit and explicit memory for new associations in normal and amnesic patients. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*, 501–518.

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.

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

Hinton, G. E. (1989). Deterministic Boltzmann learning performs steepest descent in weight-space. *Neural Computation, 1*, 143–150.

Hirsh, R. (1974). The hippocampus and contextual retrieval of information from memory: A theory. *Behavioral Biology, 12*, 421–444.

Hirsh, R. (1980). The hippocampus, conditional operations, and cognition. *Physiological Psychology, 8*, 175–183.

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.

Ikeda, J., Mori, K., Oka, S., & Watanabe, Y. (1989). A columnar arrangement of dendritic processes of entorhinal cortex neurons revealed by a monoclonal antibody. *Brain Research, 505*, 176–179.

Jacoby, L. L., Yonelinas, A. P., & Jennings, J. M. (1996). The relation between conscious and unconscious (automatic) influences: A declaration of independence. In J. D. Cohen, & J. W. Schooler (Eds.), *Scientific approaches to the question of consciousness* (pp. 13–47). Hillsdale, NJ: Lawrence Erbaum 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.

Kohonen, T. (1984). *Self-organization and associative memory.* Berlin: Springer-Verlag.

Landauer, T. K., & Dumais, S. T. (1997). a solution to plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211–240.

Lubow, R. E. (1989). *Latent inhibition and conditioned attention theory.* Cambridge: Cambridge University Press.

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. (1994). The interaction of nature and nurture in development: A parallel distributed processing perspective. In P. Bertelson, P. Eelen, & G. D'Ydewalle (Eds.), *Current advances in psychological science: Ongoing research* (pp. 57–88). Hillsdale, NJ: Erlbaum.

McClelland, J. L., & 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 connectionst models of learning and memory. *Psychological Review, 102*, 419–457.

McClelland, J. L., & Rumelhart, D. E. (Eds.). (1988). *Explorations in parallel distributed processing: a handbook of models, programs, and exercises.* Cambridge, MA: MIT Press.

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., & Barnes, C. A. (1990). From cooperative synaptic enhancement to associative memory: Bridging the abyss. *Seminars in the Neurosciences, 2*, 403–416.

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 Earlbaum Associates.

Miller, K. D., Keller, J. B., & Stryker, M. P. (1989). Ocular dominance column development: Analysis and simulation. *Science, 245*, 605–615.

Milner, B. (1966). Amnesia following operation on the temporal lobe. In C. W. M. Whitty, & O. L. Zangwill (Eds.), *Amnesia* (pp. 109–133). London: Butterworth & Co.

Minsky, M. L., & Papert, S. A. (1969). *Perceptrons.* Cambridge, MA: MIT Press.

Mishkin, M., & Petrie, H. L. (1984). Memories and habits: Some implications for the analysis of learning and retention. In L. R. Squire, & N. Butters (Eds.), *Neuropsychology of memory* (pp. 287–296). New York: Guliford Press.

Movellan, J. R. (1990). Contrastive Hebbian learning in the continuous Hopfield model. In D. S. Touretzky, G. E. Hinton, & T. J. Sejnowski (Eds.), *Proceedings of the 1989 Connectionist Models Summer School* (pp. 10–17). San Mateo, CA: Morgan Kaufman.

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

Munakata, Y., McClelland, J. L., Johnson, M. J., & Siegler, R. S. (1997). Rethinking infant knowledge: Toward an adaptive process account of successes and failures in object permanence tasks. *Psychological Review, 104*, 686–713.

Nadel, L. (1994). Multiple memory systems: what and why, and update. In D. L. Schacter, & E. Tulving (Eds.), *Memory systems 1994.* Cambridge, MA: MIT Press.

Nadel, L., & O'Keefe, J. (1974). The hippocampus in pieces and patches: an essay on modes of explanaion in physiological psychology. In R. Bellaires, & E. G. Gray (Eds.), *Essays on the nervous system. a festshrift for j. z young.* (pp. 367–390). Oxford: The Clarendon Press.

Nowlan, S. J. (1990). Maximum likelihood competitive learning. In D. S. Touretzky (Ed.), *Advances in neural information processing systems, 2* (pp. 574–582). San Mateo, CA: Morgan Kaufman.

Oja, E. (1982). A simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology, 15*, 267–273.

O'Keefe, J., & Dostrovsky, J. (1971). The hippocampus as a spatial map: Preliminary evidence from unit activity in the freely-moving rat. *Brain Research, 34*, 171–175.

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

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

O'Reilly, R. C. (1996b). *The leabra model of neural interactions and learning in the neocortex.* PhD thesis, Carnegie Mellon University, Pittsburgh, PA, USA.

O'Reilly, R. C. (1998). Six principles for biologically-based computational models of cortical cognition. *Trends in Cognitive Sciences, 2*(11), 455–462.

O'Reilly, R. C., Braver, T. S., & Cohen, J. D. (in press). A biologically based computational model of working

memory. In A. Miyake, & P. Shah (Eds.), *Models of working memory: Mechanisms of active maintenance and executive control.* New York: Cambridge University Press.

O'Reilly, R. C., & McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall: Avoiding a tradeoff. *Hippocampus, 4*(6), 661–682.

O'Reilly, R. C., & Munakata, Y. (in press). *Cognitive neuroscience: A computational exploration.* 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.* Cambridge, MA: MIT Press.

Penfield, W., & Milner, B. (1958). Memory deficits produced by bilateral lesions in the hippoaompal zone. *Archives of neurology anf Psychiatry, 79,* 475–497.

Phillips, R. G., & LeDoux, J. E. (1992). Differential contribution of amygdala and hippocampus to cued and contextual fear conditioning. *Behavioral Neuroscience, 106,* 274–285.

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.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. E. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review, 103,* 56–115.

Risold, P. Y., & Swanson, L. W. (1996). Structural evidence for functional domains in the rat hippocampus. *Science, 272,* 1484–1486.

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

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.* San Diego, CA: Academic Press.

Rolls, E. T. (1996). A theory of hippocampal function in memory. *Hippocampus, 6,* 601–620.

Rudy, J. W., & O'Reilly, R. C. (submitted). Contextual fear conditioning, conjunctive representations, and the hippocampus.

Rudy, J. W., & Sutherland, R. J. (1989). The hippocampal formation is necessary for rats to learn and remember configural discriminations. *Behavioural Brain Research, 34,* 97–109.

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. (1986a). Learning internal representations by error propagation. In D. E. Rumelhart, J. L. McClelland, & PDP Research Group (Eds.), *Parallel distributed processing. volume 1: Foundations* (Chap. 8, pp. 318–362). Cambridge, MA: MIT Press.

Rumelhart, D. E., McClelland, J. L., & PDP Research Group (Eds.). (1986b). *Parallel distributed processing. volume 1: Foundations.* Cambridge, MA: MIT Press.

Rumelhart, D. E., & Zipser, D. (1986). Feature discovery by competitive learning. In D. E. Rumelhart, J. L. McClelland, & PDP Research Group (Eds.), *Parallel distributed processing. volume 1: Foundations* (Chap. 5, pp. 151–193). Cambridge, MA: MIT Press.

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.

Schacter, D. L., & Graf, P. (1986). Preserved learning in amnesic patients: Perspectives from research on direct priming. *Journal of Clinical and Experimental Neuropsychology, 6,* 727–743.

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

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

Seress (1988). Interspecies comparison of the hippocampal formation shows increased emphasis on the regio superior in the ammon's horn of the human brain. *J Hirnforsh, 29,* 335–340.

Shapiro, M. L., & Olton, D. S. (1994). Hippocampal function and interference. In D. L. Schacter, & E. Tul-

ving (Eds.), *Memory systems 1994*. Cambridge, MA: MIT Press.

Squire, L. R. (1987). *Memory and brain*. Oxford: Oxford University Press.

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. (1994). Declarative and nondeclarative memory: Multiple brain systems supporting brain systems. In D. L. Schacter, & E. Tulving (Eds.), *Memory systems 1994*. Cambridge, MA: MIT Press.

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*. San Diego, CA: Academic Press.

Sur, M., Garraghty, P., & Roe, A. W. (1988). Experimentally induced visual projections into auditory thalamus and cortex. *Science, 242*, 1437–1441.

Sutherland, R. J., McDonald, R. J., Hill, C. R., & Rudy, J. W. (1989a). Damage to the hippocampal formation in rats selectively impairs the ability to learn cue relationships. *Behavioral and Neural Biology, 52*, 331–356.

Sutherland, R. J., McDonald, R. J., Hill, C. R., & Rudy, J. W. (1989b). Damage to the hippocampal formation in rats selectively impairs the ability to learn cue relationships. *Behavioral and Neural Biology, 52*, 331–356.

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.

Suzuki, W. A. (1996). The anatomy, physiology and functions of the perirhinal cortex. *Current Opinion in Neurobiology, 6*, 179.

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

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

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving, & W. Donaldson (Eds.), *Organization of memory* (pp. 381–403). San Diego, CA: Academic Press, Inc.

Tulving, E. (1983). *Elements of episodic memory*. Oxford: Clarendon Press.

Van Hoesen, G. W. (1982). The parahippocampal gyrus. New observations regarding its cortical connections in the monkey. *Trends in Neurosciences, 5*, 345–350.

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.

Whishaw, I. Q., & Tomie, J. A. (1991). Acquisition and retention by hippocampal rats of simple, conditional, and configural tasks using tactile and olfactory cues: Implications for hippocampal function. *Behavioral Neuroscience, 105*, 787–797.

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.

Yonelinas, A. P. (1994). Receiver-operating characteristics in recognition memory: Evidence for a dual-process model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20*, 1341–1354.

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