Psychological Review
2001, Vol. 108, No. 2, 311-345

0033-295X/01/$5.00 DOI: 10.1037//0033-295X.108.2.311

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

Randall C. O'Reilly and Jerry W. Rudy
University of Colorado at Boulder

The authors 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 is contradicted by the fact that hippocampally lesioned rats can learn nonlinear discrimination problems that require conjunctive representations. The authors' framework accommodates this finding by establishing a principled division of labor, where the cortex is responsible for slow learning that integrates over multiple experiences to extract generalities whereas the hippocampus performs rapid learning of the arbitrary contents of individual experiences. This framework suggests that tasks involving rapid, incidental conjunctive learning are better tests of hippocampal function. The authors implement this framework in a computational neural network model and show that it can account for a wide range of data in animal learning.

The role of the hippocampus in memory has been characterized in many different ways, but one common idea is 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* that binds together stimulus features specific to locations, form *episodic* memories, 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 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 encodes representations of stimulus conjunctions is critical to the following important approaches to understanding the hippocampal formation:

- Human amnesia associated with damage to the hippocampal formation 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 maplike 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 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).

All these approaches incorporate the idea that the hippocampus is important for acquiring representations of stimulus conjunctions and predict that damage to the hippocampal formation should impair performance on problems that require the acquisition of such representations. Sutherland and Rudy (1989) suggested a strong test of this prediction using nonlinear discrimination problems that can only be solved if subjects construct conjunctive representations of stimuli. This prediction resulted in a large literature that failed to support the conjunctive idea, showing instead that rats with extensive damage to the hippocampus can solve nonlinear discrimination problems that require conjunctive representations (e.g., Alvarado & Rudy, 1995b; Bunsey & Eichenbaum, 1996; Gallagher & Holland, 1992; McDonald et al., 1997; Whishaw & Tomie, 1991).

Although much of this literature has focused on disproving the specific predictions made by Sutherland and Rudy (1989), we argue that these data constitute an important challenge for many

Randall C. O'Reilly and Jerry W. Rudy, Department of Psychology, University of Colorado at Boulder.

This research was supported in part by National Science Foundation Grant IBN-9873492, National Institutes of Health (NIH) Program Project MH47566 and NIH Grant MH061316.

We thank David Huber, Yuko Munakata, Lynn Nadel, and Ken Norman for comments and discussion. Ken Norman can be credited with the general idea of understanding the transitivity results using pattern completion.

Correspondence concerning this article should be addressed to Randall C. O'Reilly, Department of Psychology, University of Colorado at Boulder, 345 UCB, Boulder, Colorado 80309. Electronic mail may be sent to oreilly@psych.colorado.edu.

other hippocampal theories that embrace the idea that the hippocampus encodes conjunctive representations. The rejection of a strong form of conjunctive theory, in our view, puts the field in a state of crisis because there is no longer a clear theoretical basis for understanding the division of labor between the hippocampus and neocortex.

In this article, we attempt to resolve this crisis by providing a theoretical framework based on two complementary but powerful learning systems, the neocortex and the hippocampus (McClelland et al., 1995). The neocortex (also called cortex) has powerful learning capacities that enable it to gradually encode regularities over many experiences. These regularities can include the contingencies of complex tasks, including the nonlinear discrimination problems that require conjunctive representations. However, there is a fundamental conflict between extracting regularities over experiences and encoding the specifics of individual experiences, such that a complementary learning system is needed in the form of the hippocampus. The hippocampal system can rapidly learn about individual experiences without suffering interference by keeping the representations of these experiences separated. Conjunctive representations emerge naturally as a result of this separation process.

Thus, we argue that stimulus conjunctions can be acquired by two neural systems, the hippocampus and neocortex. However, 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 toward 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. The slow learning of task-driven conjunctions is consistent with the way that rats actually solve nonlinear discrimination problems, thereby explaining why hippocampal lesions do not necessarily impair performance on these tasks.

Our framework suggests a class of tasks that should provide a much better test of the differential contributions of the neocortex and hippocampus than the nonlinear discrimination learning problems. Specifically, rapid, incidental conjunctive learning tasks, where the acquisition of stimulus conjunctions is not forced by task demands and only relatively few exposures are provided, should be uniquely sensitive to hippocampal damage. This is supported by a number of experimental findings (e.g., Fanselow, 1990; Hall & Honey, 1990; Honey & Good, 1993; Honey, Watt, & Good, 1998; Kim & Fanselow, 1992; Save, Poucet, Foreman, & Buhot, 1992). Our framework can also explain, at a mechanistic level, why the hippocampus appears to be important for supporting some kinds of flexibility, for example in transitive inference tasks (e.g., Bunsey & Eichenbaum, 1996; Dusek & Eichenbaum, 1997).

The article proceeds in several stages. First, we provide a historical overview of the development of the idea that the hippocampus is critical to the acquisition of conjunctive representations. We then detail how tests of Sutherland and Rudy's (1989) configural association theory generated a strong challenge to this idea and created a crisis for mechanistic accounts of hippocampal function. After reviewing another literature that is consistent with our proposed solution to this crisis, we describe the solution in detail. We then present a biologically based computational model of the hippocampal–neocortical system that instantiates our ideas about the dimensions along which the hippocampus and neocortex differ. This model is then applied to a wide range of tasks that have been used to assess the contribution the hippocampus makes to learning and memory, including nonlinear discrimination tasks, rapid incidental learning tasks, contextual fear conditioning, and transitive inference tasks.

We focus our application on animal experiments because they have most directly addressed the nature of underlying mechanisms through careful lesion studies and analytic experiments. However, the same model has also been used to account for human memory data (O'Reilly, Norman, & McClelland, 1998). Because the major aspects of our model can be motivated independently on the basis of 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

We track two themes in this overview of the historical development of theories of hippocampal function: (a) general ideas about the existence and nature of the division of labor between the cortex and hippocampus and (b) the specific idea that the hippocampus can bind together different types of information into a conjunctive representation. We track these themes through human and animal studies, and biological-computational models.

Human Studies

As is well known, the story of the hippocampus as a major contributor to human memory began about 40 years ago with the work of Milner and her colleagues (Milner, 1966; Penfield & Milner, 1958; Scoville & Milner, 1957). On the basis of 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 damage to the hippocampal formation 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 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 (Gaffan, 1974; Hirsh, 1974; Nadel & O'Keefe, 1974; see Squire, 1992, for a review).

The early, more mechanistically oriented accounts of human hippocampal function emphasized the idea that the hippocampus encodes stimulus conjunctions (Marr, 1971; Teyler & Discenna, 1986; Wickelgren, 1979). This notion continues to be central as an explanation of how people recall and recognize episodes from the past. For example, this idea was clearly embedded in the memory indexing theory of Teyler and Discenna (1986), who suggested that each experiential event is represented in a unique 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 were included in 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 that in both of these accounts the hippocampus represents the conjunction of the stimulus features that made up a particular event or experience; 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—that is, memory for the specific contents of individual episodes or events (Tulving, 1972, 1983; Tulving & Markowitsch, 1998).

In contrast with these views supporting an essentially conjunctive story, some other perspectives are more difficult to characterize in terms of the underlying mechanisms. A good example of this is the influential declarative/explicit versus nondeclarative/implicit memory distinction, which appears to provide a reasonable account of some of the differences between the hippocampal/medial-temporal lobe areas and other cortical and subcortical areas in humans (Squire, 1987, 1992). However, the lack of a clear mechanistic basis to these ideas makes them difficult to relate to the kinds of constructs that have been developed in the animal and computational literatures, which are the focus of this article.

Animal Studies

Milner's (1966) conclusion that the hippocampus plays an essential role in human amnesia also generated a large volume of animal experimental work. Although the initial findings were only indirectly related to the conjunctive learning idea, this idea soon became a dominant theme in the animal literature, although this theme took various different guises.

The first wave of studies, summarized in a thorough review by Douglas (1967), overwhelmingly demonstrated that rats and primates with extensive damage to the hippocampus and related cortical structures displayed no anterograde or retrograde amnesia for basic learning paradigms. Nevertheless, 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).

On the basis of this pattern of results, Douglas (1967) offered the hypothesis that the hippocampus was critical for enabling animals to withhold responding—the *response inhibition* view. However, Douglas realized that only certain types of responses were inhibited by the hippocampus, specifically those involving acquired stimulus–response associations. This specificity to acquired associations kept alive the possibility that the hippocampal formation was involved in memory processing in animals, even if it was in an inhibitory capacity. Also, Douglas provided the first seeds of the idea that the hippocampus plays an important role in solving the *ambiguous cue* problem. This problem emerges when the same stimulus is associated with incompatible outcomes (e.g., associated with reward in one context but not in another), and solving the problem requires keeping the resulting associations separate to minimize *associative interference.* We show later in this article that use of a separation mechanism to avoid interference is closely related to one of the functional properties of conjunctive representations.

The issue of how to solve the associative interference problem was subsequently addressed by Hirsh (1974), who proposed one of the first multiple memory system frameworks (see also Nadel & O'Keefe, 1974). Hirsh argued that a learning experience leaves its impact on two different memory systems: the *performance line* storage system and the *memory* system, which is associated specifically with the hippocampus. Generally speaking, experience leaves its effect on the performance line by altering the strength of connections between the neural elements activated by a stimulus and those 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 this is appropriate to the task at hand. In contrast, Hirsh's memory system stored representations of experience off the performance line and used the concept of a contextual label to keep conflicting associations separate. As Hirsh put 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 formerly true. (p. 426)

Constructing contextualized representations clearly involves representing the conjunctions of stimuli, behaviors, and associated outcomes as separate from these features individually—in other words, though Hirsh did not use this terminology, a conjunctive representation.

The ideas of Nadel and O'Keefe (1974) emerged most clearly 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. This system supports the acquisition of a map-like representation of the environment, where the 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 because it represents consistent rules, routes, procedures, and stimulus–response habits.

Because the hippocampus-dependent locale system represents experience as connections between stimulus features (e.g., distance, directions), it is clearly a stimulus conjunction theory. However, O'Keefe and Nadel (1978) limited the kind of information the locale system could represent exclusively to spatial infor-

mation in the form of an allocentric spatial map. Their view of the hippocampus has generated an enormous amount of research on both the physiology and memory functions of the hippocampus, and its fundamental behavioral prediction—that damage to the hippocampus will impair performance in spatial learning tasks—has been confirmed many times (cf. Barnes, 1988). However, many theorists have noted that this spatial map view is overly restrictive relative to the range of nonspatial behaviors impaired by hippocampal damage, especially in humans (Hirsh, 1980; Squire, 1992, 1994). Thus, it may be more useful to consider spatial conjunctions as a special case of a more general conjunctive hippocampal function (e.g., McClelland et al., 1995; McNaughton & Nadel, 1990; Sutherland & Rudy, 1989).

The idea that the hippocampal formation contributes to memory by representing stimulus conjunctions emerged unambiguously in an article 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 thereby represents the whole chunk" (Wickelgren, 1979, p. 44). Wickelgren's concept of chunking is clearly equivalent to the concept of conjunctive representations. The conjunctive idea 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.

Perhaps the strongest statement of the conjunctive idea came with the Sutherland and Rudy (1989) *configural association theory*, which has much in common with the ideas of Hirsh (1974) and Wickelgren (1979) reviewed earlier. The core idea in this theory was the assertion that the hippocampus is essential to the 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. This configural notion was also offered as a more general alternative to the O'Keefe and Nadel (1978) spatial map theory (Wood, Dudchenko, & Eichenbaum, 1999).

Biological/Computational Models

The anatomy and physiology of the hippocampus has been the subject of much investigation (for reviews see Amaral & Witter, 1989; Risold & Swanson, 1996; Rolls, 1989; Squire, Shimamura, & Amaral, 1989; Van Hoesen, 1982). These biological data, together with related computational neural network models, led to independently motivated theories of conjunctive encoding in the hippocampus. Two major biological properties of the hippocampus led to these ideas: (a) the considerable convergence of a wide range of different cortical areas into the hippocampus and (b) the presence of substantial interconnectivity among neurons within the CA3 region of the hippocampus.

The hippocampus receives information from virtually all association 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 (EC), the hippocampus also projects back to these areas via return connections from the EC. 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) suggested 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).

The interconnectivity among the CA3 neurons of the hippocampus figured centrally into Marr's (1971) influential computationally motivated theory of hippocampal function. Marr sought to infer the computational properties of the hippocampus from its anatomy and physiology, and he focused on the notion of 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 because it encodes a unitary representation of a stimulus pattern composed of many separable features. McNaughton and Nadel (1990) noted the similarity of Marr's concept of an auto-associator to Hebb's (1949) idea of a cell assembly and referred 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., Hasselmo & Wyble, 1997; Levy, 1989; McClelland et al., 1995; McNaughton & Nadel, 1990; O'Reilly & McClelland, 1994; Rolls, 1989).

Summary

This brief review indicates that significant aspects of the behavioral, neuroanatomical, and computational literatures have converged over the past 25 years on the idea that the hippocampal formation provides a substrate for representing stimulus conjunctions. That is, the hippocampus binds together disparate cortical representations into a unitary encoding that can later be recalled from partial cues. This idea emerged early in the history of the field, and it is at the core of many contemporary theories of hippocampal function.

Problems for Mechanistic Hippocampal Theories

Given the broad support for the importance of the hippocampus in encoding stimulus conjunctions, it is surprising that a substantial literature now seriously challenges this idea. Much of this literature was generated in response to the configural association theory of Sutherland and Rudy (1989). Perhaps Sutherland and Rudy's most important contribution is that they 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 requiring configural associations that can be solved by normal animals. 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 prototypical 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 (Minsky & Papert, 1969; Rumelhart, McClelland, & PDP Research Group, 1986). 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. 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. Thus, the solution to such a problem 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, Sutherland and Rudy (1989) 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 a simple, clear hypothesis to directly test the configural/conjunctive theory of hippocampal function.

The existing literature at that time suggested 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; this result has been replicated several times (e.g., Alvarado & Rudy, 1995b; Sutherland, McDonald, Hill, & Rudy, 1989; Sutherland et al., in press). Nevertheless, when Rudy and Sutherland (1995) reviewed additional tests of the theory, they were forced to conclude that the strong position they staked out in 1989 could not be maintained. 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 describe only two results here, with more discussion later in the context of our computational model (also see Rudy & Sutherland, 1995, for a review). 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 *AC*+, *BC*−, *AD*−, *BD*+, where each element is equally often associated with reward (+) and nonreward (−). The stimulus elements were two different diameter strings (*A* and *B*) and two odors (*C* and *D*). On a trial (e.g., *AC*+ vs. *AD*−), 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 an *ambiguous feature* 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 et al., 1997).

Many researchers agree that this literature provides ample evidence against Sutherland and Rudy's (1989) 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). However, as we noted previously, the idea that the hippocampus is specialized for encoding conjunctive representations is also central to many other theories; therefore, these data should be equally damaging to all of these theories. Nevertheless, these broader implications have not been widely acknowledged, possibly because the extent to which, at a mechanistic level, conjunctive representations are an essential component of many theories has not been sufficiently appreciated. Indeed, many theories are stated without reference to specific mechanistic constructs like conjunctive representations (e.g., the notion that the hippocampus is important for encoding declarative information or for supporting the flexible use of relational knowledge), even though we would argue that conjunctive representations provide an essential mechanism for such ideas.

Once the central importance of conjunctive representations as a mechanistic principle is appreciated, however, it is clear that the findings of preserved conjunctive learning under hippocampal damage have implications that extend beyond Sutherland and Rudy's (1989) conjunctive theory. If the function of the hippocampus cannot be identified with a clear mechanistic principle, such as enabling the learning of conjunctive representations, then what is the alternative, other than ad hoc descriptions of data or vague amechanistic terminology? Furthermore, how can these descriptive ideas be related to the highly specialized neural structure of the hippocampal formation? Either an alternative mechanism needs to be put forth or the idea that the hippocampus stores representations of stimulus conjunctions must be constrained in a way that places theorizing about the hippocampus on rational ground.

Other Conjunctive Tasks: Hints of a Way Out

Nonlinear discrimination problems unambiguously require the subject to learn conjunctive representations. Indeed, they cannot be solved unless the requisite conjunctions are learned. Conjunctive representations, however, can also be learned even when they are not required to solve any problem. The tasks used to study this incidental conjunctive learning are quite simple. Subjects are exposed to a set of features in a particular configuration and then the features are rearranged. Subjects are then tested to determine if they can detect the rearrangement. If the test indicates that the rearrangement was detected, then one can infer the subject learned a conjunctive representation of the original configuration. The literature indicates that the incidental learning of stimulus conjunctions, unlike many nonlinear discrimination problems, is dependent on the hippocampus. After reviewing this literature, we integrate it with the nonlinear discrimination literature to show how together they are consistent with a principled understanding of the division of labor between the cortex and the hippocampus that is the basis for our theoretical framework.

Rapid, incidental conjunctive learning in animals. Perhaps the simplest demonstration 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 encoded representations of the individual objects and could discriminate them from novel objects. However, only the control rats encoded the conjunctions necessary to represent the spatial arrangement of the objects, even though this was not in any way a requirement of the task.

A more recent article by Honey et al. (1998) makes a similar point. They repeatedly exposed control rats and rats with excitotoxic hippocampal lesions 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 a train of clicks was followed by a flashing light on the right side. After the orienting response to the constant and flashing light in both sets of rats habituated, the auditory and visual combinations were switched (the clicks preceded the constant light and the tone signaled the flashing light). This switch reinstated the orienting response to the light 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. 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 of the *context specificity* effect that normal rats, but not rats with hippocampal damage, learn stimulus conjunctions that are not required by the task (Good & Bannerman, 1997; Hall & Honey, 1990; Honey & Good, 1993; Honey, Willis, & Hall, 1990). In these studies, rats are conditioned to cue *A* in Context 1 and cue *B* in Context 2, and then they are tested in switched contexts (cue *A* in Context 2 and cue *B* in Context 1). Normal rats, but not those with hippocampal damage, exhibit more conditioning in the original contexts than in the switched ones. Because each of the contexts and stimuli were equally associated with reward, responses based on the independent elements should not exhibit this context specificity effect (Rudy & Sutherland, 1995). Thus, the intact rats were incidentally encoding conjunctions between the context and stimulus elements whereas the hippocampally lesioned ones were not.

Evidence for the involvement of the hippocampal formation in the incidental learning of stimulus conjunctions has also emerged in the contextual fear conditioning literature. 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 (e.g., a tone) paired with shock (Kim & Fanselow, 1992; Phillips & LeDoux, 1992, 1994; but see Maren, Aharonov, & Fanselow, 1997). Fanselow (1990; see also Kiernan & Westbrook, 1993) argued that hippocampally mediated contextual fear conditioning derives from conjunctive representations of context on the basis of the following data. If intact 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 hr later. However, they do show fear if they are in the conditioning chamber for about 2 min before being shocked. Fanselow argued that this additional time was necessary for the construction of a conjunctive representation of the conditioning context before the shock occurred. Consistent with this interpretation, Fanselow showed that 2 min of exposure to the conditioning context 24 hr prior to immediate shock resulted in contextual fear conditioning. He argued that this 2-min exposure was sufficient to permit the animals to (incidentally) construct a configural, unitary representation of context, which was then associated with fear during the subsequent immediate shock.

In summary, there are conditions under which animals automatically acquire representations of stimulus conjunctions as a natural consequence of being exposed to the environment. The examples cited here 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 idea that the hippocampus, but not the cortex, naturally develops conjunctive representations. One salient source of 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 a different context, performance was better (by roughly 15%) 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 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 nonamnesic participants were helped by such cues. Control and amnesic participants were matched for performance on recognizing the words without context, so the lack of facilitation in amnesic patients cannot be attributed to a floor effect. Further evidence comes from a recent study by Chun and Phelps (1999), in which specific context facilitated visual search for intact participants but not hippocampally damaged patients. 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 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 that, together with the properties of an event, defines a

particular episode (Tulving, 1972). Because such an episode is generally unique, it must be learned rapidly as the episode unfolds. Further, the contextual information is typically incidental to any task that might happen to be performed at the time, yet such information appears to be encoded automatically. There is evidence that episodic recall (but not necessarily recognition, though this is somewhat controversial) is specifically impaired in patients with selective hippocampal damage (Holdstock et al., in press; Vargha-Khadem et al., 1997).

Finally, it is likely that the rapid conjunctive learning supported by the hippocampus operates in many situations used to test people in which task demands do not force such learning. For example, consider a set of simple, linearly solvable discrimination learning problems (e.g., $A+$ vs. $B-$; $C+$ vs. $D-$; $E+$ vs. $F-$). Such problems could be solved either by rapid conjunctive learning of the cue and consequent outcomes as supported by the hippocampus or by gradual incremental learning supported by the cortex. Neurologically intact people solve such problems in very few trials, whereas patients with damage to the hippocampus solve them more gradually (Reed & Squire, 1999; Squire, Zola-Morgan, & Chen, 1988). Such data can thus be viewed as reflecting the rapid conjunctive learning available to intact people but not to patients with selective damage to the hippocampus. However, these data do not directly implicate the use of conjunctive representations—tests in which the elements of the original task rearranged in novel combinations are required to assess conjunctivity (as in the animal studies described previously).

Summary: Two Types of Conjunctive Learning

There is a potentially conflicting and confusing pattern of hippocampal dependence across the nonlinear discrimination and incidental conjunctive learning tasks, even though all these tasks involve conjunctive representations. To clarify this pattern, it is important to discriminate between two types of conjunctive learning. One type is associated with nonlinear discrimination problems, where conjunctive learning emerges in the service of problem solving and requires a substantial amount of training. The other type is associated with incidental tasks, where conjunctive learning occurs rapidly and automatically. In the next section, we show that computational neural network principles of learning in the cortex and hippocampus clearly predict that the hippocampus should be important for the incidental tasks but not necessarily the nonlinear discrimination problems. This analysis provides a way out of the theoretical crisis.

A Complementary Cortical/Hippocampal Memory System Framework

At the center of our framework is a set of principles for understanding how the cortex functions. It is clear that the cortex is important for many of the most important aspects of preserved learning after hippocampal damage (though many other areas, such as the basal ganglia, amygdala, and cerebellum, also play important roles, e.g., Davis, 1992; Fiez, 1996; Gao, Parsons, & Fox, 1996; LeDoux, 1992; Mishkin, Malamut, & Bachevalier, 1984; Packard, Hirsh, & White, 1989). For example, damage to cortical areas surrounding the hippocampus impairs several aspects of learning that are spared with more selective hippocampal lesions. Our principles of cortical functioning, based on a variety of considerations at the biological, psychological, and computational levels of analysis, clearly support the idea that the cortex is capable of powerful learning.

Nevertheless, our model of cortical learning also has important limitations: It cannot rapidly acquire representations of novel experiences. This limitation indicates a fundamental tradeoff between learning the general features of an environment and learning the specifics of a particular experience (McClelland et al., 1995; Sherry & Schacter, 1987). The cortex is specialized for gradually extracting generalities, and the hippocampus is specialized for rapidly learning the specifics that define a particular experience. Although our model assumes that the cortex and hippocampus constitute two complimentary learning systems, we think that both operate according to a common set of underlying mechanistic principles. Their unique contributions are a product of key differences in their architecture and other parameters, including the overall level of activity (*sparseness*) and the learning rate.

We begin by describing the core cortical principles and then discuss their limitations and how the hippocampus can provide complementary learning functions. We then discuss in more detail how a few central features of the hippocampal system can lead to its unique learning capacities. We conclude with a summary of the critical differences between the cortex and the hippocampus, and how in general these account for the empirical data presented previously. A number of important issues raised by our framework are discussed next, followed by our explicit computational model that implements our theoretical ideas and demonstrates their ability to account for a wide range of data.

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, such as object recognition, spatial processing, language, working memory, planning, and so on. Furthermore, the cortex is generally regarded as a highly plastic system capable of extensive experience-dependent learning. Putting these views together, it is reasonable to conclude that the cortex is a highly capable system even in the absence of the hippocampal system (though there are other views on this, as we discuss later). Here, we provide a set of arguments centered around computational neural network modeling principles to support and elaborate this idea.

Computational neural network models have been developed that use learning mechanisms to understand human language, perception, and other high-level cognitive abilities. These models are typically based on either error-driven *backpropagation* learning (Rumelhart, Hinton, & Williams, 1986) or on statistically based self-organizing learning mechanisms that utilize Hebbian-like mechanisms (e.g., Miller, Keller, & Stryker, 1989). We incorporate both of these learning mechanisms in our model (O'Reilly & Munakata, 2000; O'Reilly, 1996b, 1998). With these two mechanisms, the cortex can be modified by task demands (by error-driven learning) and can represent the extent to which different features co-occur (by Hebbian learning). Together, these learning mechanisms enable the cortex to extract the invariant properties of repeated experience but not the unique features of each experience.

After elaborating our model of cortical learning, we then explore some ramifications of this model in the next section.

Error-driven task learning. The backpropagation mechanism for performing error-driven learning minimizes errors in performance by iteratively adjusting the weights between connected units in the direction that will most decrease the error. Critically, this mechanism can also modify connectivity between hidden layers of units interposed between input and output units. Because hidden units in the cortex can be modified to represent stimulus conjunctions, the cortex should in principle be able to solve nonlinear discriminations without assistance from the hippocampus.

However, backpropagation has been widely challenged on the grounds that it lacks a plausible biological mechanism (e.g., Crick, 1989; Zipser & Andersen, 1988). Specifically, backpropagation requires that an error value is propagated backwards from the dendrite of a receiving neuron, across the synapse, into the axon terminal of the sending neuron, down the axon of this neuron, then integrated and multiplied by some kind of derivative, and then propagated back out of its dendrites. Moreover, no one has ever recorded anything that resembles an error signal.

However, a well-documented property of the cortex, bidirectional connectivity, can be used to perform essentially the same error-driven learning as backpropagation (O'Reilly, 1996a). 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. Another oft-cited problem with backpropagation concerns the origin of the teaching patterns that provide the error signals. However, many potential sources for these teaching patterns in the form of actual environmental outcomes 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.

Hebbian model learning. Use of Hebbian learning mechanisms to represent co-occurrence (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 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 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.

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, in press; O'Reilly & Munakata, 2000). 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, 2000). In addition, Hebbian learning can explain phenomena such as preserved repetition priming in persons with amnesia (e.g., Schacter & Graf, 1986), where there are no obvious sources of error or task demands to drive the learning.

Limitations of Cortical Learning and the Need for Complementary Systems

Although we believe that the model described in the preceding section 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 catastrophic levels of interference when applied to a list learning task. Although many attempts were made to remedy this failure, McClelland et al. (1995) concluded that this failure reflects a fundamental tradeoff in learning. On the one hand, successful adaptation requires organisms to extract and represent the general properties of the environment. On the other hand, it 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, and so on.

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, incremental learning that integrates over specific instances, whereas acquiring specifics often requires fast learning that keeps the specific instances separate. The requirement that integrative learning be slow for neural network learning mechanisms was proved by White (1989) and is discussed further in McClelland et al. (1995). The basic intuition is captured by the idea that the weights connecting units in a network represent a kind of running average over experiences, and the time window over which any kind of running average is computed is directly proportional to the size of the time constant (learning rate), with smaller (slower) values giving longer time windows of integration.

To avoid the fundamental tradeoff between learning about generalities versus specifics, it is reasonable that the brain would use two complementary learning and memory systems that optimize these objectives separately. 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. This computationally motivated division of labor between cortex and hippocampus is generally consistent with other descriptive characterizations (e.g., O'Keefe & Nadel, 1978; Sherry & Schacter, 1987) and other models (e.g., Alvarez & Squire, 1994; Hasselmo & Wyble, 1997). In particular, Sherry and

Schacter (1987) suggested an almost identical distinction between learning invariances across episodes versus learning the variances of particular episodes, with the further suggestion that incompatible functions such as these provide an important criterion for distinguishing between memory systems.

The nature of this hippocampal/cortical tradeoff could also be mapped onto the semantic versus episodic distinction advocated by Mishkin, Vargha-Khadem, and Gadian (1998) and Tulving and Markowitsch (1998), in that semantic memory typically refers to knowledge about the general nature of the world. However, depending on one's definition of the term *semantic memory,* such memories can also include rapidly acquired specific information that would involve hippocampal learning. Thus, we prefer to use mechanistically explicit terminology regarding the contributions of the hippocampus and cortex.

Principles of Hippocampal Function

On the basis of the preceding discussion, to complement the cortex, the hippocampus should rapidly acquire information about a specific experience and represent it so that interference produced by its similarity to other experiences is minimized (e.g., where you parked your car today vs. yesterday). In this section, we build on the framework developed for understanding the cortex to provide a set of principles of hippocampal function and show how certain architectural and parametric properties of the hippocampus can support rapid conjunctive learning while minimizing interference. To reduce interference produced by overlapping input patterns, the hippocampus supports *pattern separation* by using a relatively small number of highly selective units to represent an input pattern (i.e., a *sparse* representation). This also produces conjunctive representations. A complete memory system, however, not only must store input patterns but it must also permit their retrieval. For the hippocampus to support memory retrieval, it must be capable of performing *pattern completion,* where a subset of cues from a previous experience can activate (retrieve) the stored pattern representing that experience. Thus, the hippocampal architecture and operating parameters must balance two countervailing functions, pattern separation and pattern completion. These mechanisms are described in more detail below.

Pattern separation. Both pattern separation and conjunctive representations are produced when an input pattern is represented by a small number of active neural units. To understand why a sparse representation can lead to these outcomes, consider a situation where the hippocampal representation is generated at random with some fixed probability of a unit becoming active. In this case, if fewer units are active, the odds decrease that the same units will be active in two different patterns (Figure 1). For example, if the probability of becoming active for one pattern (i.e., the sparseness) is .25, then the probability of becoming active for both patterns would be $.25^2$ or .0625. If the patterns are made more sparse so that the probability becomes .05 for being active in one pattern, the probability of being active in both patterns falls to .0025. Thus, the pattern overlap is reduced by a factor of 25 by reducing the sparseness by a factor of 5 in this case. However, this analysis does not capture the entire story because it fails to take into account the fact that hippocampal units are actually driven by weighted connections with the input patterns and therefore will be affected by similarity (overlap) in the input.

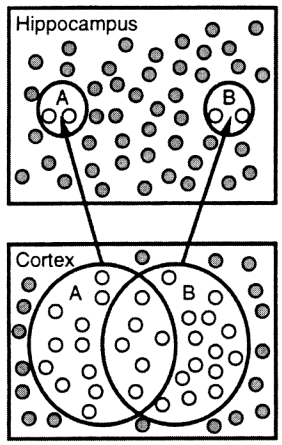

Figure 1. Pattern separation in the hippocampus. Small circles represent units, with active ones in white and inactive ones in gray. Circles A and B in the cortex and hippocampus indicate two sets of representations composed of patterns of active units. In the cortex, they are overlapping and encompass a 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.

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 becomes active. In reality, multiple units (roughly 1–5%) are activated in the hippocampus at any given time, but the same principle applies (see O'Reilly & McClelland, 1994, for a detailed analysis).

For optimal pattern separation, it is important that different receiving units be maximally activated by different input patterns. This can be achieved by having relatively diffuse, random patterns of partial connectivity with the inputs, which appears to be a property of the perforant path of the hippocampus (as discussed in

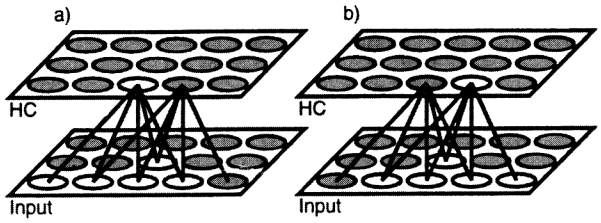

Figure 2. Conjunctive, pattern-separated representations result from sparseness (active units are represented in white, inactive ones in gray). The extreme case where only one receiving unit (in the upper layer, representing the hippocampus) is allowed to be active is shown here for simplicity. Each 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 yield completely nonoverlapping (separated) memory representations. In this way, the conjunctive memory representation resulting from sparseness produces pattern separation. HC = hippocampus.

greater detail later). One important consequence of this random conjunctivity is that it suggests that the hippocampus acts as a simple *binding* device (Cohen & O'Reilly, 1996) instead of forming more systematic "relational" encodings (e.g., Eichenbaum, 1992), which would seem to require more systematic patterns of connectivity. Under this simple binding view, all relationship information must be present in the inputs to the hippocampus, which can then bind together the relational information with other information about the related items in a conjunction. For example, the cortex would encode that the chair is to the left of the table (left-of being the relational encoding), but the hippocampus could bind this information together with details about the specific properties of the chair and table into a unitary representation.

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 (as emphasized in Marr's, 1971, auto-associator theory).

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 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; in the other, it has to produce an 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 (e.g., your friend) require a judgment call involving a tradeoff between pattern separation and completion.

In addition to providing basic recall of stored information, pattern completion can enable some kinds of flexible processing that the cortical system by itself cannot support. This flexibility arises by pattern completing to stored memories based on novel input cues. In short, although the cortex can perform some degree of both pattern separation and completion, the unique features of the hippocampal system (principally sparse representations and extensive auto-associator circuitry) produce much more significant capacities for these important functions.

Complexities of the separation/completion tradeoff. The fact that pattern separation and completion trade off with each other is important for understanding the behavior of the hippocampus in nonlinear discrimination tasks. The critical dimension for determining whether pattern separation or pattern completion will occur in the hippocampus is the overlap (similarity) of the input patterns. Figure 3, based on a simulation from O'Reilly and McClelland (1994), summarizes the separation/completion tradeoff as a function of the level of input pattern overlap—for very high levels of overlap, pattern completion takes over from pattern separation. Usually, such high levels of overlap would only be present in cases where the input is a retrieval cue for a previously stored pattern. When a large number of features contribute to the input pattern, as for representations of environmental context, even ostensibly similar inputs, such as two different views of the environment, will likely have enough differences to drive pattern separation, not completion. However, many nonlinear discrimination learning problems prove to be an important exception to this rule because they specifically recombine a small number of stimulus elements across conditions that require conflicting outputs.

In these nonlinear discrimination problems, the hippocampus can be using pattern completion to recall previously stored patterns

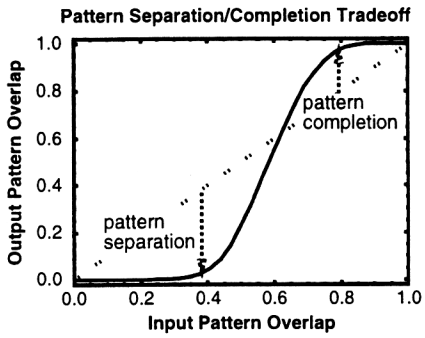

Figure 3. Tradeoff between pattern separation and completion as a function of input overlap (similarity) between two random patterns. The vertical axis shows the overlap in the hippocampal (simulated rat-sized CA3) representation of input patterns having the level of overlap specified on the horizontal axis. The diagonal line shows the identity transformation—values below this line reflect pattern separation and those above the line reflect pattern completion. As similarity increases, pattern completion takes over from separation. Details of the learning mechanism can alter where this tradeoff line falls, but its existence is a basic property of the network. From "Hippocampal Conjunctive Encoding, Storage, and Recall: Avoiding a Tradeoff," by R. C. O'Reilly and J. R. McClelland, 1994, *Hippocampus, 4,* p. 674, Figure 15a. Copyright 1994 by Wiley & Sons, Inc. Adapted with permission.

in situations where pattern separation would otherwise be more advantageous. When this occurs, error-driven learning, operating within the hippocampus in much the same way it operates in the cortex, can overcome the pattern completion process to produce pattern separation, but in this case it will likely take many repetitions of learning. Thus, in these situations, learning in the hippocampus can look a lot like that of the cortex, as we see when we apply our computational model to the nonlinear discrimination learning problems.

Principled Account of Conjunctive Learning

We now describe how this theoretical framework can, in principle, provide an account of performance on tasks that require the learning of conjunctive representations. To summarize, the critical properties for understanding cortical and hippocampal differences are

> *Learning rate.* The cortical system typically learns slowly, whereas the hippocampal system typically learns rapidly.
>
> *Conjunctive bias.* The cortical system has a bias toward integrating over specific instances to extract generalities. The hippocampal system is biased by its intrinsic sparseness to develop conjunctive representations of specific instances of environmental inputs. However, this conjunctive bias trades off with the countervailing process of pattern completion, so the hippocampus does not always develop new conjunctive representations (sometimes it completes to existing ones).
>
> *Learning mechanisms.* Both cortex and hippocampus use error-driven and Hebbian learning. The error-driven aspect responds to task demands and will cause the network to learn to represent whatever is needed to achieve goals or ends. Thus, the cortex can overcome its bias and develop specific, conjunctive representations if the task demands require this. Also, error-driven learning can shift the hippocampus from performing pattern separation to performing pattern completion, or vice versa, as dictated by the task. Hebbian learning operates constantly 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. There are two key findings: (a) the cortex alone can learn nonlinear discrimination problems; and (b) the hippocampus, but not the cortex, is capable of rapidly forming conjunctive representations in incidental learning contexts.

The finding that the cortex will develop conjunctive representations over a relatively large number of trials when such representations are specifically required by the task (e.g., to obtain rewards) is entirely consistent with the idea that error-driven learning is operating in the cortex. This kind of learning is specifically driven by task contingencies and can form complex conjunctive representations when given enough training trials. As we discussed previously, the hippocampus also requires many repetitions of error-driven learning to learn some of these nonlinear tasks because it ends up performing pattern completion instead of pattern separation.

Therefore, an important conclusion from our framework is that, ironically, nonlinear discrimination problems do not reveal the unique contributions of the hippocampus precisely because they require that the subject develop conjunctive representations. These tasks are learned slowly and they cannot be solved unless the subject develops representations of stimulus conjunctions. The cortex can acquire conjunctions under these conditions. Instead, our framework suggests that incidental learning tasks that do not require the subject to learn stimulus conjunctions provide the best way to reveal the contributions of the hippocampus. The critical feature of such tasks is that the subject rapidly acquires representations of stimulus conjunctions even though they are not required by any task demands.

The second major conclusion from our framework, therefore, is that these rapid, incidental learning tasks provide the best venue for assessing the role of the hippocampus in learning.

A Computational Neural Network Model

We now describe a computational model that implements our theoretical framework. The model is based on a computational framework called *Leabra* (O'Reilly, 1996b, 1998, in press; O'Reilly & Munakata, 2000), which 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 use of a common underlying set of mechanisms is supported by the numerous structural similarities between cortex and hippocampus (which is a form of cortex called archicortex), including many of the same general patterns of interconnectivity between excitatory pyramidal neurons and inhibitory neurons and the same kinds of synaptic modification (i.e., learning) mechanisms. After we briefly summarize the basic network mechanisms, we discuss 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 the model's behavior.

Basic Mechanisms

The equations for these mechanisms are presented in the Appendix, and the main properties are 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 the two basic mechanisms discussed earlier: 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

(Kohonen, 1984; Nowlan, 1990; Rumelhart & Zipser, 1986). 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.

The overall architecture and connectivity of the model is shown in Figure 4. 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 are described in greater detail in the next section. The hippocampus interfaces with the cortex via the entorhinal cortex (EC), which captures the information represented in the cortex in a one-to-one fashion. 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 a 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.

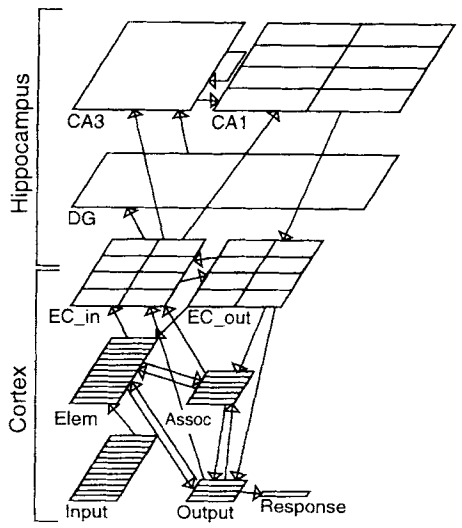

Figure 4. The model, showing both cortical and hippocampal components. The cortex has 12 different input dimensions (sensory pathways), with four different values per dimension. These are represented separately in the elemental cortex (Elem). Higher level association cortex (Assoc) can form conjunctive representations of these elements, if demanded by the task. The interface to the hippocampus is via the entorhinal cortex (EC), 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. *DG* = dentate gyrus.

Although we have attached different labels to the cortical and hippocampal components, they are really both part of the same bidirectionally connected network. Activity simultaneously flows between the cortical and hippocampal parts at each step of updating; the development of cortical representations can affect the trajectory of hippocampal learning and vice versa. This results in complex interactions that can be difficult to analyze in detail, but the model nevertheless captures the overall contributions of the cortex and hippocampus that our theoretical framework suggests.

The Cortical System

All of the representations in the cortical system are organized into groups of four units (shown in Figure 4 as the smaller boxes within the cortical layers), with only one out of these four 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, while providing 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 four units that we refer to as a *slot,* representing four different values along each dimension, and there are a total of 12 such dimensions (slots). 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 develops distributed representations over six four-unit slots. 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 is thought to correspond to the parahippocampal region in the rat.

Although it typically only represents a binary reward/no-reward value, the output layer also has a population-coded representation over four slots. This distributed output representation is important for providing a sufficiently substantial representation of the output layer in the hippocampal system, relative to the other cortical areas (which all contribute several active units to the hippocampal input). The output layer 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 layer 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, 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.

After the information is encoded 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. Then 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 recall from taking place.

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.

The model similarly 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

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		Model	
Area	Neurons	Activity %	Units	Activity %
EC	200,000	7.0	96	25.0
DG	1,000,000	0.5	250	1.6
CA3	160,000	2.5	160	6.3
CA1	250,000	2.5	256	9.4

Note. EC = entorhinal cortex; DG = dentate gyrus. Rat data are from Barnes, McNaughton, Mizumori, Leonard, and Lin (1990); Boss, Peterson, and Cowan (1985); Boss, Turlejski, Stanfield, and Cowan (1987); and Squire et al. (1989).

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, which 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, CA1 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 nonsystematic 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 eight such columns. Each column receives input from three 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 three `EC_in` slots (64 different combinations) in a corresponding set of three `EC_out` slots. Thus, each CA1 has a conjunctive, pattern-separated representation of the patterns within the three 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.

To capture the idea that the hippocampus learns incidentally and automatically, 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 see when we apply the model to nonlinear discrimination problems. Also, the learning rate is twice as fast in the hippocampus compared with the cortex (.02 vs. .01) to facilitate its rapid learning. This cortical learning rate is the standard value for most complex, interleaved learning problems in Leabra (O'Reilly & Munakata, 2000). That the hippocampal rate is only twice as fast suggests that the specialized features of the hippocampal anatomy also play an important role in producing rapid learning effects.

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.

In our simulations, we focus on the qualitative, not quantitative, features of the data. This is because, with only the slight modifications needed to accommodate a few of the more complex experimental paradigms, we use exactly the same model for all of our simulations. To produce more detailed quantitative fits, we would expect that various parameters would need to be tuned to reflect the different details present across different experiments, which would undermine our main point, that a single set of principles can account for the critical (qualitative) patterns across a wide range of behavioral data.

Nonlinear Discrimination Problems

The primary goal of these simulations is to show that our model can solve nonlinear discrimination problems without the contribution of the hippocampal component. Our theoretical framework emphasizes that this cortical conjunctive learning arises from the explicit task demands of these problems—these task demands are captured by the error signals that drive learning in both the cortical and hippocampal components of our model. Also, we argue that these problems trigger hippocampal pattern completion instead of pattern separation, such that even the intact animal takes many trials to learn them. Beyond these basic points, more complex patterns of data exist in the literature that suggest that some nonlinear problems are more sensitive to hippocampal damage than others. Although these patterns are not completely reliable across studies, our model reproduces what appears to be the dominant pattern.

Negative patterning, ambiguous feature, and biconditional problems. We begin by analyzing three problems: (a) the negative patterning (NP) problem, $A+$, $B+$, $AB-$; (b) the ambiguous feature (AF) problem, $AC+$, $B+$, $AB-$, $C-$, studied by Gallagher and Holland (1992); and (c) a version of the biconditional discrimination, $CA+$, $CB-$, $DA-$, $DB+$. First we compare the very similar NP and AF problems. Both of these problems require many trials to learn, even for intact subjects, and rats with hippocampal damage are able to learn them with enough trials. Nevertheless, there are a number of reports that rats with damage to the hippocampal formation are impaired relative to intact control rats on the NP problem (e.g., Alvarado & Rudy, 1995b; McDonald et al.,

1997; Rudy & Sutherland, 1995) but not on the AF problem (Alvarado & Rudy, 1995b; Gallagher & Holland, 1992). Indeed, in spite of their similarity, Alvarado and Rudy (1995b) reported that the same animals that were impaired on NP were not impaired on AF. However, Davidson et al. (1993) found no impairment for hippocampal lesions on the NP problem, so there may also be other relevant task factors or individual differences at work here.

The NP and AF problems were implemented in the model by presenting the patterns shown in Figure 5. Note that, following Alvarado and Rudy (1995b), we added the $C-$ trial to the NP problem, making it even more similar to AF 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 AF 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 vs. interleaved training, which we explore later).

Figure 6 compares the performance of the intact and lesioned models on the NP and AF problems with the data from Alvarado and Rudy (1995b). These comparison data make four points: (a) of most importance, both the intact and lesioned models can solve these problems; (b) both problems require many trials to solve (both models make many errors); (c) the intact model performs no better on the AF problem than the lesioned model; but (d) consistent with the bulk of the literature, the intact model is better than the lesioned model on the NP problem. Thus the model's behavior closely matches the data.

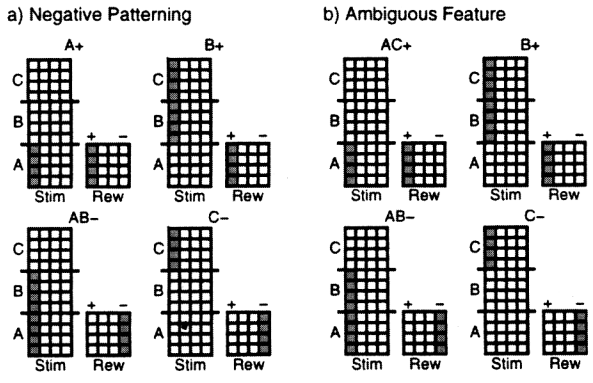

Figure 5. Input/output patterns for the (a) negative patterning and (b) ambiguous feature problems. For each of the four trial types in each problem, the input stimuli (Stim) and output reward (Rew) are shown. Mutually exclusive values (e.g., + vs. − reward) are represented as different values within a dimension, whereas 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.

McDonald et al. (1997) examined the role of the hippocampal formation in several nonlinear discriminations, including the NP problem and a biconditional problem. Both problems required many trials to solve, and they presented evidence that rats with damage to the hippocampus acquired the stimulus conjunctions. In addition, however, rats with damage to the hippocampus were more impaired on the NP problem than they were on the biconditional problem. In fact, depending on whether one looks at the transformed or nontransformed data from their experiment, damage to the hippocampus either had no effect or a modest effect (see also Whishaw & Tomie, 1991).

The stimulus elements in the McDonald et al. (1997) 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% 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).

Figure 7 shows the patterns we used to implement the biconditional. As shown in Figure 8, consistent with the literature indicating that rats with damage to the hippocampus solve the biconditional problem, the lesioned model performed as well as the intact model. This problem was also difficult and the models required many trials to solve it.

Explanation of the model's behavior. The network produces the two most basic findings from the literature: (a) The cortex alone can solve nonlinear discrimination problems, and (b) these problems are difficult and require many trials to be solved. The first outcome can be explained as the result of error-driven learning shaping the units in association cortex to construct the conjunctive representations needed to solve the problem. Consistent with this interpretation, the cortical model could not solve any nonlinear discrimination problems if either the association cortex units were removed or the error-driven learning process was not used.

Nonlinear discrimination problems require many trials to solve, even in the intact model, because of the tradeoff between the pattern separation and pattern completion properties of the hippocampus (see Figure 3). These problems require pattern-separated conjunctive representations of the controlling stimuli, but because there is extensive overlap in the input patterns that have to be conjoined the pattern completion properties of the hippocampus are engaged. Pattern completion then interferes with the need to associate different outcomes with these similar patterns.

For example, solving the NP problem ($A+, B+, AB-$) requires that the animal construct a representation of the *AB* compound that is separated from the representations of *A* and *B*. However, when *A* or *B* is presented, the hippocampus will have a strong tendency to pattern complete to the *AB* representation. In such cases, the *AB* representation, in addition to the *A* or *B* representation, would

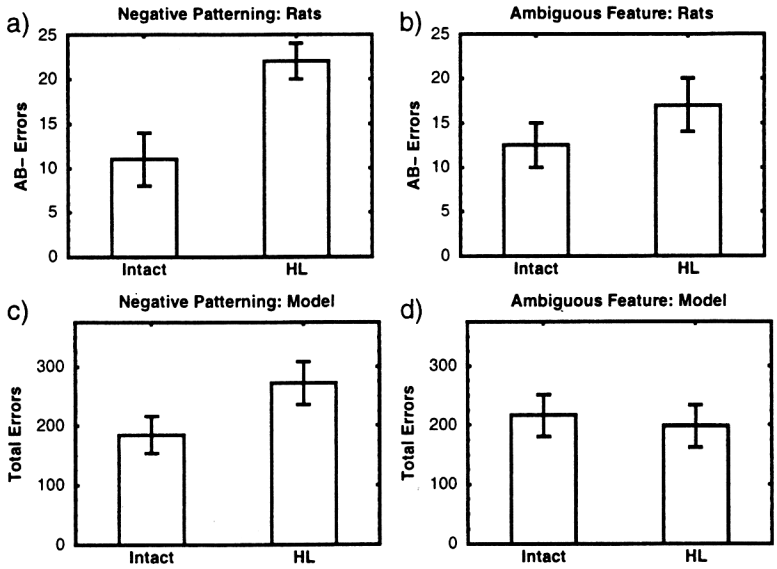

Figure 6. Results for the negative patterning (left column) and ambiguous feature (right column) problems. The top row shows data from rats from Alvarado and Rudy (1995b), and the bottom row shows data from the model. Intact is intact rats/networks, and HL is rats/networks with hippocampal lesions. $N = 40$ different random initializations for the model. The hippocampally lesioned system is able to learn the problems, and all conditions require many trials (i.e., large number of errors). Negative patterning is differentially impaired with a hippocampal lesion.

become more associated with reward, which works against the solution to the problem.

Our model also captures the pattern in the literature indicating that the NP problem depends more on the hippocampus than do the AF or biconditional problems. In approaching this outcome, it is important to appreciate that the difference between the intact and lesioned models' performance on these problems is small compared with the number of trials needed to solve them. Also, damage to the hippocampus also does not always impair performance on the NP problem (e.g., Davidson et al., 1993).

The specific difficulty with the NP problem has actually already been identified by Gallagher and Holland (1992) and Rudy and Sutherland (1995), who noted that the extent to which the individ-

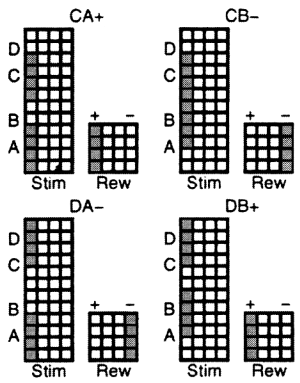

Figure 7. Input/output patterns for the biconditional discrimination problem studied by McDonald et al. (1997), where *A* and *B* stimuli overlap 50%, as do *C* and *D*. Stim = stimuli; Rew = reward.

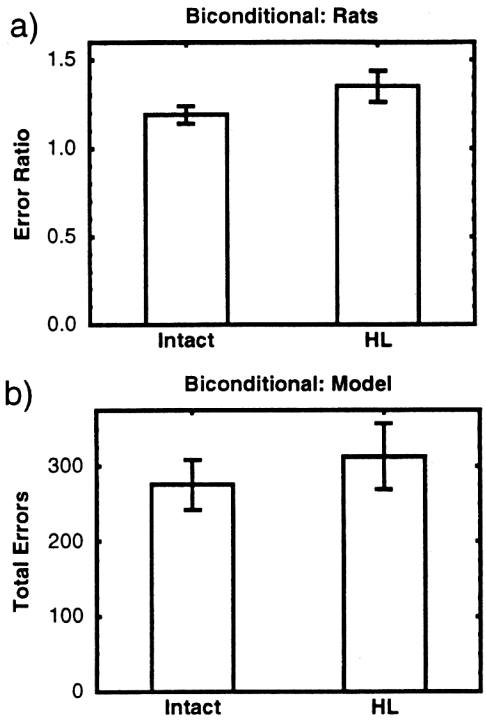

Figure 8. Results for the biconditional problem for (a) rats (data from McDonald et al., 1997, replotted in terms of error ratios) and (b) the model, which shows no statistically reliable difference between the intact and lesioned conditions. HL = rats/networks with hippocampal lesions.

ual stimulus elements (e.g., *A, B, C*) appear alone versus in combination with other elements was an important difference between NP and AF (and the biconditional). In the NP problem, both *A* and *B* (and *C*) appear alone, whereas in the AF problem, only *B* (and *C*) appear alone. In the biconditional problem, no elements appear alone. We think this difference is important because it has implications for the relative difficulty the network (and the animal) has in separating individual elements appearing alone (e.g., separating *A* and *B* from *AB* in the NP problem) as compared with separating combinations of elements (e.g., separating *AC* from *AB* in the AF problem).

The problem with elements appearing alone is that it is very difficult to form a conjunction with only one stimulus input, yet these conjunctions are essential for separating the representations in nonlinear problems (Figure 9). However, this problem is present for both the cortex and the hippocampus, so why is the hippocampus of any benefit? We answer this question using a "horse race" analogy. The cortex and the hippocampus are both attempting to separate the elements (*A, B*) from the compound (*AB*) in the NP problem and both systems require many trials. The hippocampus may have a slight advantage in this race because its sparse representations, compared with the cortex, make it somewhat easier to allocate different, nonoverlapping subsets of units to represent the elements and the compound. The sparseness advantage of the hippocampus is less important when the elements appear in compounds, as in the AF and biconditional problems, and is also countered by the greater tendency of the hippocampus to pattern complete.

The next section provides further support for this analysis by probing the extent to which the internal representations of the *A* element and *AB* compound are truly separated in the NP and AF problems.

Assessment of pattern separation and blocked versus interleaved training. Our analysis suggests that the NP problem requires that the representation of the *A* element be separated from the *AB* compound, whereas the AF problem can rely on the interactions with the *C* stimulus to separate *AC* from *AB* (as illustrated in Figure 9). Alvarado and Rudy (1995a) provided evidence relevant to this issue. They trained one set of intact rats to solve the AF 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 significantly increased errors on *AB* trials for rats previously trained on the AF problem but had no effect on the errors made by rats trained on the NP problem, exactly as our analysis would suggest.

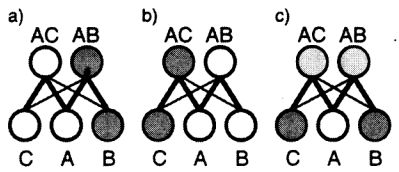

Figure 9. Example of how the presence of multiple stimuli enables the network to easily represent conjunctions. Lighter units are more active. If *A* is seen in the presence of *C, AC* is favored, and in the presence of *B, AB* is favored, but if just *A* is present, there is nothing to modify or interact with, so all representations that have an *A* in them (*AC* and *AB* in this case) are equally activated. Thus, negative patterning is specifically difficult because it has two out of three trials where the stimulus elements appear alone.

We simulated the Alvarado and Rudy (1995a) experiment in our model and found the same results. As shown in Figure 10, additional *A*+ training increased the number of errors on the *AB*− trials made by rats trained on the AF problem compared with rats trained on the NP problem. To further support our analysis that the reason the cortex has greater difficulty on the NP problem is because it has greater difficulty separating *A* from *AB*, we found that the lesioned network exhibited 10.6 *AB*− errors on this test compared with only 2.4 for the intact network.

Alvarado and Rudy (1995a) also compared two versions of the NP problem. In one case rats were trained in a standard way: All trial types (*A*+, *B*+ and *AB*−) were pseudorandomly interspersed in each session. In another case, the rats received blocked presentations of the trial types, with *A*+ trials presented in one block, and *B*+ and *AB*− trials in another. These rats were then given *A*+ trials and tested on the interleaved NP problem as described previously. Rats in the blocked condition increased their errors (responses) on *AB*− compounds compared with 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 10).

We can explain these results 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 remapping 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.

On the basis of this analysis, we expected that our lesioned model would not be impaired on the blocked version of the NP problem because, unlike the interleaved NP problem, the blocked version does not force the model to construct pattern-separated representations of *A, B,* and *AB*. As shown in Figure 11, which compares the simulation of the blocked and interleaved problems, the lesioned and intact models did not differ on the blocked problem but did differ on the interleaved problem. Furthermore, we observed that the lesioned model had a slightly slower remapping of the response at the beginning of each block compared with the intact model, which is due to the slower learning rate in the

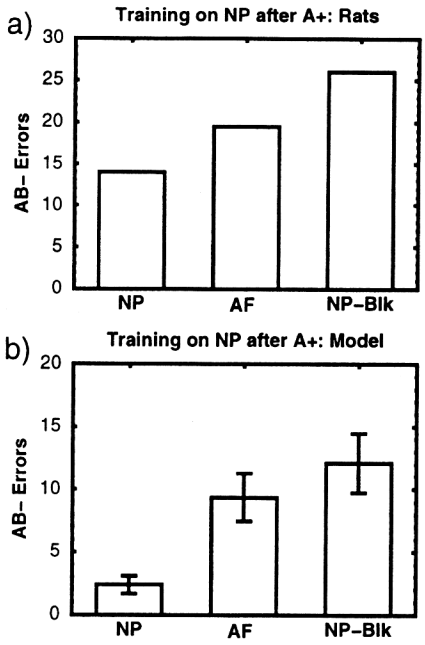

Figure 10. Results for *AB* errors in the negative patterning (NP) problem after $A+$ trials for (a) rats (data from Alvarado & Rudy, 1995a) and (b) the model. The interference from the $A+$ trials is the least in the interleaved NP problem relative to the other problem types (ambiguous feature [AF] 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. Blk = blocked.

cortex compared with the hippocampus. This produced the small difference in overall errors between the intact and lesioned models. We also expect to find these small differences in lesioned and intact rats.

Transverse patterning. Damage to the hippocampal formation impairs performance on another nonlinear discrimination problem, the transverse patterning (TP) problem (Alvarado & Rudy, 1992, 1995b, 1995c; Dusek & Eichenbaum, 1998; but see Bussey, Warburton, Aggleton, & Muir, 1999, for contrary results from fornix lesions). At first glance, this result appears to violate the explanation of why the NP problem is more dependent on the hippocampus than are the AF 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 TP requires the subject to make a choice between two stimulus elements. Specifically, the animal has to concurrently solve three simultaneous discrimination problems constructed from only three elements. Representing the correct choice as + and the incorrect choice as −, we can describe the problems as follows: $A+$ versus $B-$; $B+$ versus $C-$, and $C+$ versus $A-$. 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, 1995b, 1995c) or could be 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 its 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 NP problem, where stimuli appear individually. Thus, conjunctive representations must be constructed largely from single stimuli in the TP problem, and the sparseness of the hippocampus can make a measurable contribution.

The typical training regime for TP in rats involves three phases. First, they learn the $A+$ versus $B-$ problem, then the $B+$ versus $C-$ problem is introduced, and finally the third problem ($C+$ versus $A-$) 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, 1995b, 1995c; Dusek & Eichenbaum, 1998).

We implemented TP in the model in a manner similar to the previous problems. As shown in Figure 12, 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 $A+$ versus $B-$ trial). We used three units in the input space to represent each of the stimuli in the initial configuration (e.g., *AB*) and three units to represent the choice made (e.g., *A*). Thus, as compared with 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 TP problem (i.e., all three trial types interleaved) and just the second phase with only two of the three trial types. As shown in Figure 13, the model captures the

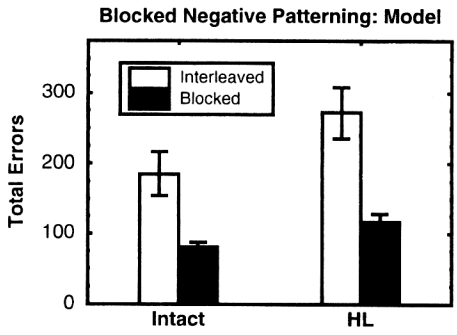

Figure 11. The model results are for learning performance in the blocked version of negative patterning for both the intact model and the model with the hippocampal component removed (HL), as compared with the standard interleaved intact and HL data presented earlier. Note that the interleaved data are taken from Figure 6.

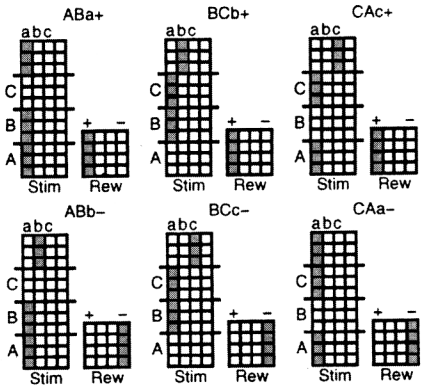

Figure 12. Input/output patterns for the transverse patterning problem. The first set of stimuli (*A–C*) represents the initial configuration prior to choice, and the second set (a–c) represents which choice was made, with the reward being based on whether the correct choice was made. Stim = stimuli; Rew = reward.

pattern of results reported in the literature, with the hippocampal lesion condition impairing performance on the full problem but not on the second phase of the problem alone (which is relatively easy for both the intact and lesioned model; 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.

Summary. Like the literature, our model shows that under some conditions the hippocampus can make a contribution to solving nonlinear discrimination problems. However, it is important to appreciate that with or without an intact hippocampus, both animals and our model require many trials to solve these problems. This is because there is extensive overlap in the stimulus patterns that have to be associated with different outcomes, and the necessary conjunctive learning is driven by the reinforcement contingencies of the tasks. The extensive overlap coupled with conflicting outcomes associated with the elements in effect neutralizes the contribution of the hippocampus to conjunctive learning. Thus, by this analysis, nonlinear discrimination tasks are not well suited to reveal the unique contribution that the hippocampus can make in encoding conjunctions.

Rapid Incidental Conjunctive Learning

We argued earlier that the hippocampal formation makes its most important contribution to memory by automatically and rapidly storing incidental stimulus conjunctions. Rapid incidental conjunctive learning 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 Honey and Good (1993) provided evidence of hippocampal-formation involvement in incidental learning by studying the context specificity of conditioning. They conditioned rats to cue *A* in Context 1 (*C*1) and Cue *B* in Context 2 (*C*2). Normal rats not only conditioned to the two cues, but they also incidently learned where the cues occurred because responding to the cues was disrupted if Cue *A* was tested in *C*2 and Cue *B* was tested in *C*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.

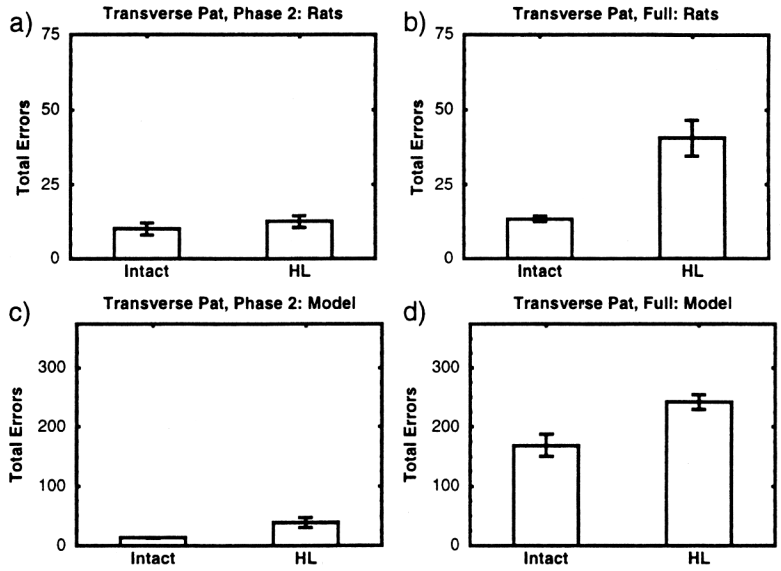

Figure 13. Results for the transverse patterning (Pat) problem, for both Phase 2 (left column), where only two out of the three trial types are used, and the full problem (right column), with all three trial types. Only the full problem requires separated conjunctive representations, and it shows an effect of hippocampal lesion (HL) relative to the intact case in both rats (top row, data from Alvarado & Rudy, 1995b) and the model (bottom row). 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.

We applied the intact and hippocampally lesioned models to the context specificity effect to see if it would simulate Honey and Good's (1993) findings. Instead of using exactly the same experimental design as Honey and Good, we used a design where the reward value of the contexts was specifically neutralized. Thus, we trained the network on two different simple discrimination problems in two different contexts: *C*1: *A*+, *B*−; *C*2: *C*+, *D*−. Because the contexts have no net reward value in our design, subjects could simply ignore the context and learn on the basis of just the individual stimuli. However, if the hippocampus is automatically encoding stimulus conjunctions, then a test where the context–stimulus pairs are switched (i.e., *C*2: *A*+, *B*−; *C*1: *C*+, *D*−) should reveal any contribution from such conjunctive representations. In Honey and Good's design, the contexts could possibly attain at least some reward value, producing a positive response bias. Indeed, the simulation results produce a clearer effect than Honey and Good's experiment, so we consider them to be a prediction for future experimental testing.

To test the model, we ran two conditions following training with either the intact or lesioned models: (a) The cues were presented in their original context 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 when the cues are tested in their original training contexts.

The specific patterns we used to train the network are shown in Figure 14. In this and all subsequent simulations, the data are based on 25 replications with random initial weights. Figure 15 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. This matches the somewhat weaker effects, indicated only by a significant interaction between lesion and test condition, seen in Honey and Good's (1993) data.

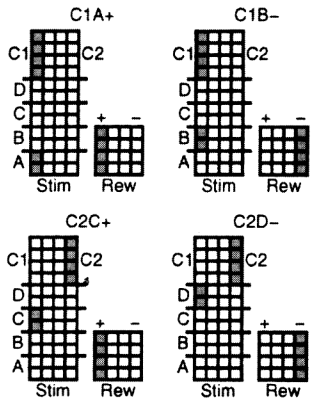

Figure 14. Input/output patterns for the incidental learning context specificity effect. Note cues *A* and *B* have equally associative values and that the two contexts C1 and C2 have no net association with reward. If rats respond only to the linear combination of context and cue associative values, then responding should be the same regardless of the context in which the cues are presented. Stim = stimuli; Rew = reward.

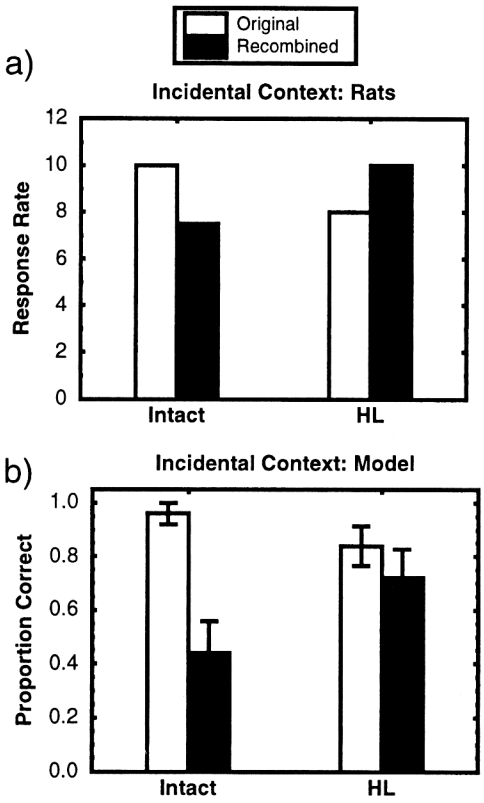

Figure 15. Incidental conjunctive learning results for testing with both the original (training) and recombined (switched) contexts. (a) Results from Honey and Good (1993; response rate). (b) Results from the model on a similar (but not identical) task (proportion correct). Even though the contexts are completely incidental to the task, the intact rats and model suffer from a context switch, whereas the rats and model without the hippocampus (HL) do not.

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 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. Because fear conditioning can be considered a simple spatial context learning task, the results here should also generalize to other spatial learning

tasks (though additional navigational mechanisms would likely also be required).

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.

Three phases of a contextual fear conditioning experiment 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; 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 16). 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 toward zero 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.

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

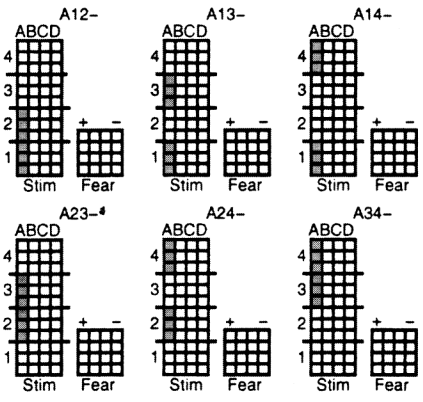

Figure 16. Input/output patterns for the exposure phase of contextual fear conditioning. All possible pairwise combinations of the four context features for the *A* environment are experienced, enabling the hippocampus to encode a conjunctive representation of the fear conditioning context. Stim = stimuli.

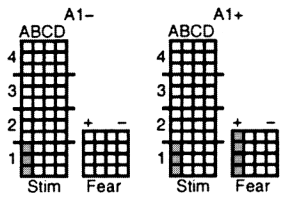

Figure 17. Input/output patterns for the shock phase of contextual fear conditioning. The + output represents a fear response induced by the shock. The input stimulus (Stim) is assumed to be a single context feature, which is arbitrarily chosen to be the first feature. The fact that the rat views the environment for a brief period prior to being shocked is represented by the initial trial without the fear output activated.

when shocked (Figure 17). Nevertheless, the intact model can pattern complete this single input to the entire context representation, which can then become associated with shock. 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 nonfear output pattern, which then interferes with the ability of the network to learn the shock-induced fear association during the shock phase. Thus, without suppressing 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

[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, 1999), and where it has been reported, a considerable amount of preexposure was necessary (Kiernan & Westbrook, 1993). Therefore, the reported results are for complete suppression 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.

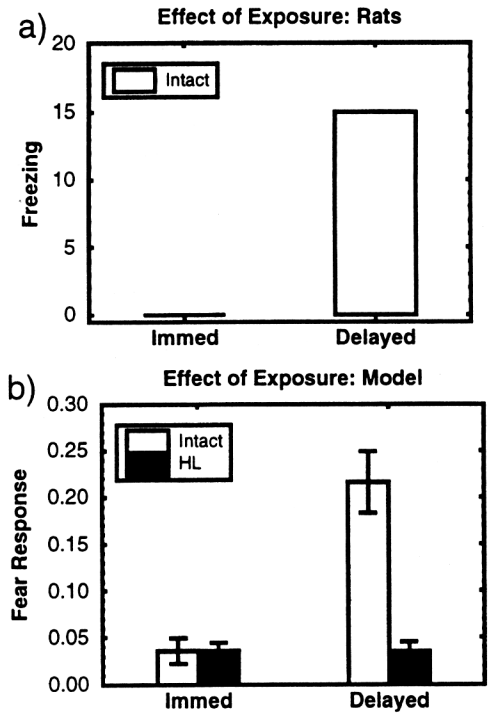

Figure 18. Effects of exposure to the context on level of fear response (a) in rats (measured by freezing, data from Fanselow, 1986) and (b) in the model, measured as the activation level for fear output units minus the baseline measure of fear response activation without any conditioning. The immediate 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 2 min (rats) or 100 epochs (model) of exposure in the environment prior to the shock, resulting in substantial conditioning in the intact rats/model, but not in the network with a hippocampal lesion (HL; no equivalent rat data available).

was to add a negative bias to the appropriate EC units so they would be inactive during exposure.

The first set of simulations demonstrates 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 18, 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.

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, although 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, 1999). In one experiment, we compared the effects of preexposure with 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 four different environments (for 100 epochs each), where each such environment had one of the four conditioning context features (Figure 19). The results of this simulation are shown in Figure 20, 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 (1999) experiment, there was a pronounced facilitation of contextual conditioning when the intact model was exposed to the context as compared with exposure to the features separately. The hippocampally lesioned network showed 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 encoded 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

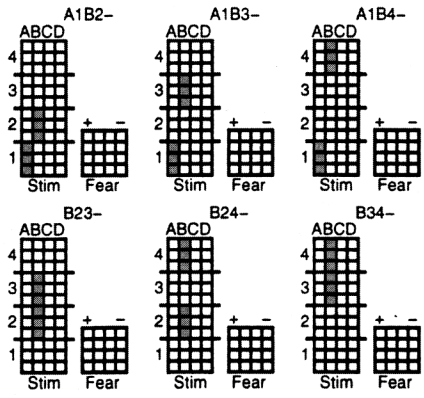

Figure 19. Input/output patterns for exposure to the conditioning context features separately. The first feature of the conditioning context (A1) is mixed in with other features defining a separate environment where this feature was experienced (B2–4). The second conditioning context feature (A2) was similarly experienced in another different environment (C2–4), and so on. Stim = stimuli.

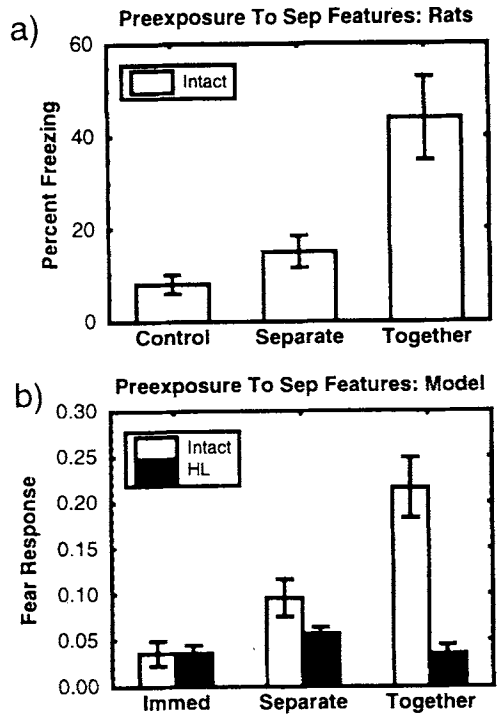

Figure 20. Effects of exposure to the features separately compared with exposure to the entire context on level of fear response in (a) rats (data from Rudy & O'Reilly, 1999) and (b) the model (see Figure 18 for details). The immediate shock condition (Immed) is included as a control condition for the model. Intact rats and the intact model show a significant effect of being exposed to the entire context together compared with the features separately, whereas the hippocampally lesioned (HL) model exhibits slightly more responding in the separate (Sep) condition, possibly because of the greater overall number of training trials in this case.

testing cues to reactivate the conjunctive context representation and its association with the shock. Recently, we provided novel evidence for the pattern completion process by studying generalized contextual fear conditioning (Rudy & O'Reilly, 1999). In this section, we show that our model replicates these pattern completion findings.

Rudy and O'Reilly (1999) constructed two contexts, *A* and *B*, which shared several features, and a Context *C* that shared no features 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 Contexts *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 preexposed 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, preexposure 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 two out of the four features) and a Context *C* that 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 21, which match those of Rudy and O'Reilly (1999). Preexposure to *A* and conditioning on *B* produced an equivalent level of fear when tested on either *A* or *B*, but preexposure to *C* yielded 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 was some level of fear response to *A* even when preexposed to *C*, we conclude that the network was also pattern completing somewhat to *B* in the *A* testing environment. The lesioned network exhibited a low level of

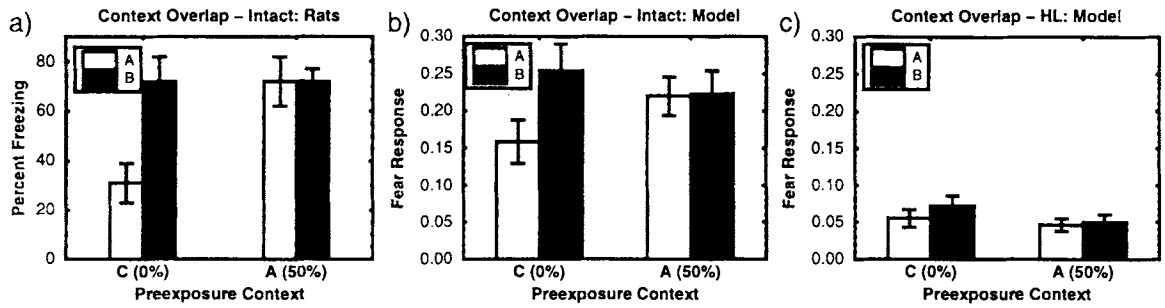

Figure 21. Effects of preexposure to contexts that overlap with the conditioning context (*B*) by an amount indicated in the horizontal axis (*A* has 50% overlap, *C* has 0% overlap). Testing performed in both *A* and *B* contexts. (a) The intact rat behavior (data from Rudy & O'Reilly, 1999) network, (b) the intact model, and (c) the hippocampally lesioned (HL) network. Pattern completion is indicated in the intact rat/model because the amount of conditioning to *A* was similar to that shown for *B* (because of 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 model did not show any differentiable effects.

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.

Summary. 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 contribution 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 vs. conjunctive feature preexposure and pattern overlap conditions as described above).

Transitivity and Flexibility

Several theorists have described memories encoded by the hippocampus as being flexible, meaning that (a) such memories can be applied inferentially in novel situations (Eichenbaum, 1992; O'Keefe & Nadel, 1978) or (b) that they are available to multiple response systems (Squire, 1992). Although the term *flexibility* provides a useful description of certain behaviors, it does not provide a mechanistic understanding of how this flexibility arises from the properties of the hippocampus. In this section, we show how the basic mechanism of hippocampal pattern completion can explain some of these flexibility phenomena while making specific testable predictions.

Some of the best evidence for hippocampal flexibility comes from studies of *transitivity* in animals (Bunsey & Eichenbaum, 1996; Dusek & Eichenbaum, 1997). In one set of problems, Dusek and Eichenbaum trained rats to solve a set of concurrent odor discriminations that took the form $A+$ versus $B-$, $B+$ versus $C-$, $C+$ versus $D-$, and $D+$ versus $E-$. Following training to criterion on these problems, rats were then given probe trials with B versus D and A versus E. When confronted with the A versus 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, the B versus 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 Bunsey and Eichenbaum's (1996) version of the transitivity test, rats were trained on two sets of conditional odor discrimination problems (Figure 22). 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 with the choice now between C and Z. Normal rats chose C when the sample was A and Z when the sample was X. Rats with damage to the hippocampal system, however, chose randomly.

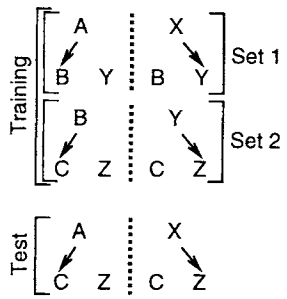

Figure 22. Logic of Bunsey and Eichenbaum's (1996) version of the transitivity test.

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. Dusek and Eichenbaum's (1997) version argued that the rats apply a transitivity operation to the B versus D case and infer that because $B > C$ and $C > D$, that it must be that $B > D$. Specifically, Dusek and Eichenbaum proposed that their rats had stored the problems as an orderly hierarchy that included all five elements of the four problems ($A > B > C > D > E$) that could be used flexibly in the service of supporting logical inferences. Similar arguments were made in Bunsey and Eichenbaum's (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, 1997, and AX, CZ in Bunsey & Eichenbaum, 1996) 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 in the following sections 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. Dusek and Eichenbaum trained the rats in ordered trial blocks, starting with 10 trials on the $A+$ versus $B-$ problem always followed by 10 trials on the $B+$ versus $C-$ problem, always followed by 10 trials on the $C+$ versus $D-$ 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 $A > B > C > D > E$ 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.

As Figure 23 shows, the overlapping hippocampal representations can then activate the correct *B* response for the *B* versus *D* probe by means of pattern completion. Specifically, if the hippocampal representations for *B*+ versus *C*− (*BC*) and *C*+ versus *D*− (*CD*) overlap, then the overlapping portion of these representations will be activated by both *B* and *D* in the *B* versus *D* probe. Because of pattern completion, one of the two hippocampal representations will be activated (*BC* or *CD*) and will produce the corresponding response (*B* or *C*, respectively). However, because *C* is not available as a choice option on the *B* versus *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* versus *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 with the preexisting bidirectional connectivity between them), which we assume the rat would naturally do. Then we trained the network in a sequential, blocked manner on 10 trials of each of the problems in order. Figure 24 shows the patterns used. We repeated this sequence five times, by which point the model had learned all the problems, and then ran 10 epochs of randomly interleaved training on all problem types. This simulates the blocked training used by Dusek and Eichenbaum (1997), except that they used successively fewer trials per block in their repetitions.

We find in the model that the final random-order training is useful to prevent a kind of *recency effect* from the blocked training. In general, the network is more likely to pattern complete the test probe to a training pattern that was more recently trained, and in the blocked training sequence, the *C*+ versus *D*− problem always follows the *B*+ versus *C*− one, and is thus more recent. Therefore, the network is more likely to pattern complete to *C*+ versus *D*−

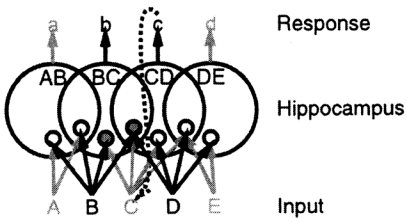

Figure 23. Illustration of how overlapping hippocampal representations can lead to correct transitivity response for the *B* versus *D* probe. The large circles each represent the collection of hippocampal units encoding a given comparison, as labeled (e.g., *AB* is *A*+ vs. *B*−). The overlap in representations 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* versus *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* versus *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*, thus making the correct response *B*.

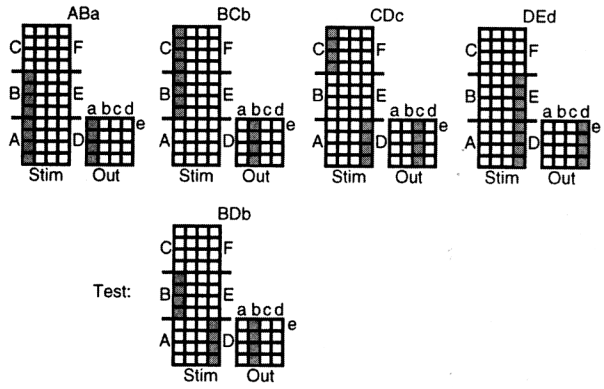

Figure 24. Input/output patterns for the $A > B > C > D > E$ version of the transitivity test (Dusek & Eichenbaum, 1997). The network is trained to produce the appropriate choice response (labeled with lowercase letters) given an input representation of the two stimuli. The top row shows the training patterns, while the bottom shows the *B* versus *D* test pattern, with the appropriate *B* response indicated in the output (which was used only to compare with the network's output). Stim = stimuli; Out = output.

instead of *B*+ versus *C*−, which increases the probability of producing the wrong output (*C*). The final interleaved training reduces this recency effect by providing recent training on all the patterns, and thus facilitates the production of the correct (*B*) output.

The results of the model are shown in Figure 25. To interpret these results, we first need to take into account an important difference between the rat and our model—the rat is forced to either choose *B* or *D*, but the model can produce any of the four trained outputs (*A* through *D*). Our model provides a good fit to the data if one assumes that the forced-choice constraint on the rat causes it to always choose *B* even when its hippocampus might have pattern completed to the *C* output by way of activating the hippocampal representation for the *C*+ versus *D*− problem. Although the intact model has some tendency to do this remapping of an initial *C* response to a *B* output (because it responds *B* about twice as often as *C*), the forced-choice constraints on the rat probably make it more likely to do so. Note that the hippocampally lesioned model has a much reduced tendency to produce the correct responses. Indeed, it seems to produce each of the four trained responses about ¼ of the time—in other words, at random.

An interesting prediction falls out of our model that would seem to directly contradict the prediction that a logical reasoning account of transitivity performance would make. This prediction concerns what would happen if one additional comparison was trained, *E*+ versus *F*−, and then the transitivity test was *B* versus *E* instead of *B* versus *D*. Logically, *B* and *E* are even further apart from each other, and thus it should be easier to conclude that *B* beats *E* than it would be to conclude that *B* beats *D*. However, according to our pattern-completion account, which depends on pattern overlap as explained previously, the fact that the hippocampal representations for *B* and *E* are further separated from each other should make it much less likely that the network will get the *B* versus *E* problem right.

The results from the model, shown in Figure 26, confirm our reasoning about the pattern-completion-based mechanism—the intact network never produces the correct response (*B*) to the *B*

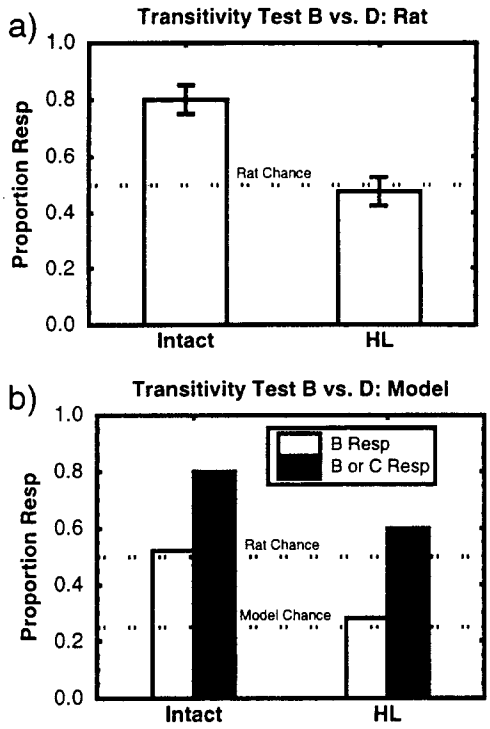

Figure 25. Results for the transitivity test (*B* vs. *D*). (a) Data from Dusek and Eichenbaum (1997) in rats. (b) Model data, showing both proportion of correct *B* responses (Resp) and proportion of either *B* or *C* responses. Because the model is not constrained to choose either *B* or *D* but the rat is, the *B* or *C* response may provide a better approximation to the rat behavior assuming that the rat does a better job of remapping initial *C* responses to actual *B* choices. The model chance line reflects the ¼ chance in the model, and the rat chance line reflects the ½ chance for the rat; the intact model's *B* responses are well above model chance, and its *B* and *C* responses are well above rat chance, whereas the hippocampally lesioned (HL) model's responses are at chance for both cases.

versus *E* probe! We also tested this network on the *B* versus *D* probe to make sure that the additional training problem was not behind the model's poor performance. These results were very similar to those shown before, ruling out this alternative account for the impaired performance on the *B* versus *E* probe. The lesioned network appears to still be responding essentially randomly (with five trained responses, chance is 20%). This prediction from the model thus stands as an important test of the two different accounts of how rats solve the transitivity problem.

The $A \rightarrow B$, $X \rightarrow Y$, . . . transitivity problem. Our analysis of Bunsey and Eichenbaum's (1996) version of the transitivity problem also depends on hippocampal pattern completion. We show that this pattern completion effect depends on the training order to produce the correct response, as a result of a recency effect. By this, we mean that a more recently experienced memory will be more frequently recalled than one that was not experienced as recently, as was mentioned previously in the discussion of the other transitivity problem.

In this case, the transitivity test probes ($A \rightarrow C$ and $X \rightarrow Z$) each overlap with two different training patterns (e.g., $A \rightarrow B$ and $B \rightarrow C$ for the $A \rightarrow C$ probe). Thus, we would expect that the hippocampus would pattern complete the $A \rightarrow C$ probe to either the $A \rightarrow B$ or $B \rightarrow C$ training representations, but not to the $X \rightarrow Y$ or $Y \rightarrow Z$ patterns, which it does not overlap with. As in the previous task, only one of these two training patterns is associated with the correct transitivity probe response ($B \rightarrow C$), so the key to solving the problem is to favor pattern completion to this training pattern (and to $Y \rightarrow Z$ for the $X \rightarrow Z$ probe). The training procedure used in this task does exactly that, by taking advantage of the recency effect.

Bunsey and Eichenbaum (1996) trained rats sequentially on the two sets of problems (Figure 22). First, they trained on $A \rightarrow B$ and $X \rightarrow Y$ until rats were performing to criterion. Then they trained on $B \rightarrow C$ and $Y \rightarrow Z$ to criterion. It was at this point that the transitivity test was given. Thus, because the training patterns having the correct responses for transitivity were trained last, these were more likely to be pattern completed to by the hippocampus (because of the recency effect), producing correct transitivity behavior.

To solve the conditional discrimination problems in this task, the rats had to maintain the sample in memory for it to conditionalize the choice. Because our model currently does not include a process to hold the sample in memory, we had to make two decisions to implement the problem within the input/output framework of the model. First we decided to model the task at the point where the choice is made and the learning occurs. Thus, the input pattern was the sample (e.g., *A*) and the choice stimulus last visited (e.g., *B*), and the output pattern was the choice response (e.g., *B*). During the early stages of training before they learned the conditionalizing pattern, rats presumably visited both wells given each sample stimulus (e.g., visiting *B* and *Y* with the *A* sample). However, as the rats mastered the problem, the incorrect well visits would drop out (e.g., visiting *Y* with the *A* sample). Thus, our second decision was to only model the stage of training where the correct $A \rightarrow B$ and $X \rightarrow Y$ choices were made, which simplified the implementation to the point where we could use our standard model (Figure 27).

We explored a representative range of three different training conditions to test our hypothesis that the recency effect of training on $B \rightarrow C$ and $Y \rightarrow Z$ was important for achieving transitivity

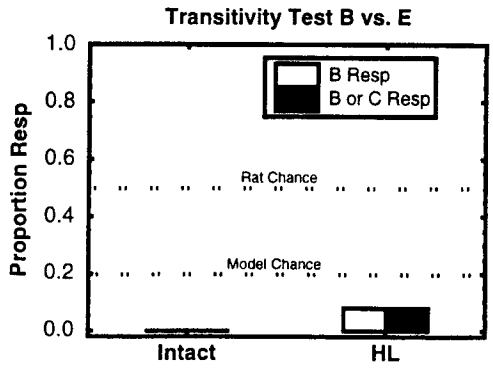

Figure 26. Model results for the *B* versus *E* transitivity test, showing both proportion of correct *B* responses (Resp) and proportion of either *B* or *C* responses. Model chance is now ⅓ instead of the ¼ shown in Figure 24. The intact network does not respond correctly at all in this case, but the hippocampally lesioned (HL) network performs somewhat near chance.

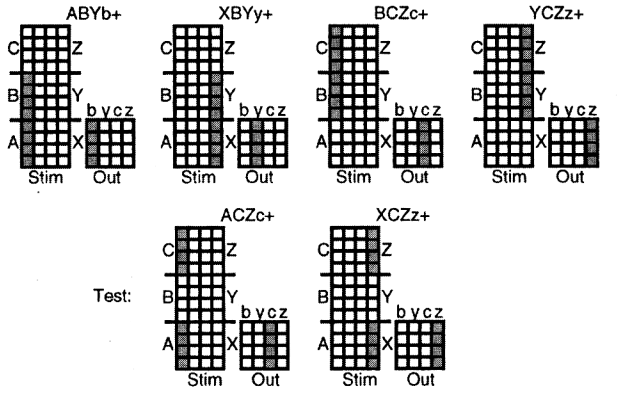

Figure 27. Input/output patterns for the $A \rightarrow X$, $B \rightarrow Y$, $C \rightarrow Z$ version of the transitivity test (Bunsey & Eichenbaum, 1996). We assume that the rat remembers the sample stimulus (Stim) and learns to make a response to the correct choice odor. Thus, in the $A \rightarrow B$ case, we represent the A and B odors in the input and train the network to produce the B response (denoted in lowercase in the figure). As before, the top row consists of the training cases, and the bottom consists of the testing cases. Stim = stimuli; Out = output.

behavior. The first condition mirrored the procedure used in Bunsey and Eichenbaum (1996), where there were two sequential blocks of training, the first on the $A \rightarrow B$ and $X \rightarrow Y$ trials and the second on the $B \rightarrow C$ and $Y \rightarrow Z$ trials. We trained for 50 trials in each block, which was sufficient to achieve mastery of the problem. The second condition still used a blocked design, but there were now 10 blocks of 10 trials alternating between the two trial types. Thus, the $B \rightarrow C$ and $Y \rightarrow Z$ trials were still the most recently trained, but the recency effect should be smaller. The final condition was randomly interleaved training on all trial types.

The results are shown in Figure 28. First, only the intact model exhibited any evidence of transitivity—the hippocampally lesioned network always performed at or below chance. Second, the importance of the recency effect in causing the network to pattern complete to the appropriate hippocampal representation is evident as a function of the training conditions: Perfect transitivity behavior is exhibited in the sequentially blocked condition (2 blocks of 50), intermediate behavior for the more fine-grained blocking (10 blocks of 10), and nonsignificantly above-chance behavior for the fully interleaved condition.

To summarize our exploration of transitivity, we have shown that hippocampal pattern completion can produce the correct transitivity responses in two different types of problems. This pattern-completion-based mechanism depends critically on the training parameters (order of training and the use of blocked training trials). Thus, an important contribution of our model is to highlight the importance of these "incidental" aspects of the experimental paradigm for achieving the transitivity outcome—these features should not be important under the "logical inference" account proposed by Eichenbaum and colleagues, but are demonstrably important in our mechanistic, pattern-completion-based account. Thus, to the extent that further empirical work finds that these training parameters are important for the rat's correct performance as well, this would constitute an important source of support for our account. Furthermore, we have highlighted the model's prediction regarding the B versus E transitivity test, which also constitutes an important test of our model.

This emphasis on the task parameters and the importance of mechanistic, process-based models is reminiscent of the general point emphasized by Munakata (1998) and Munakata, McClelland, Johnson, and Siegler (1997) that detailed task parameters can be understood in a mechanistic, neural-network-based framework in ways that simply do not make sense under more abstract symbolic-level or richly interpreted accounts.

General Discussion

The idea that the hippocampal formation contributes to memory by enabling organisms to store representations of stimulus conjunctions is central to a number of theories, and there is considerable evidence consistent with this view. However, this idea alone cannot be correct because there is direct evidence that rats with damage to the hippocampal formation can solve nonlinear discrimination problems that require conjunctive representations. The major goals of this article are to

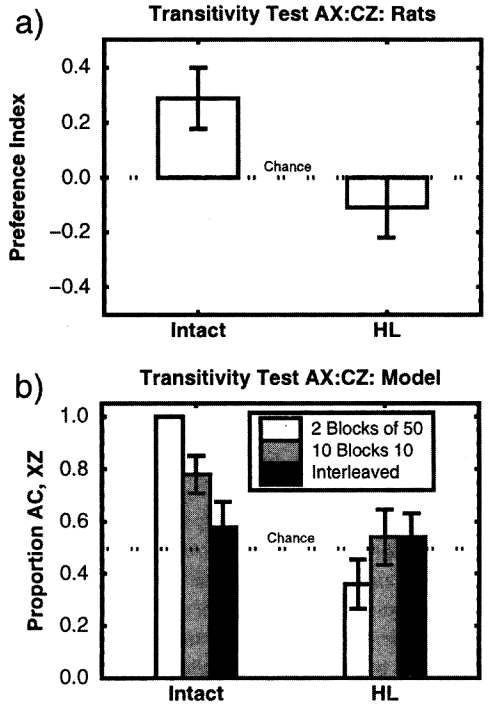

Figure 28. Results for Bunsey and Eichenbaum's (1996) transitivity test ($A \rightarrow C$ and $X \rightarrow Z$, where the transitivity-appropriate ["correct"] responses are C and Z, respectively) for (a) rats (data from Bunsey & Eichenbaum, 1996) and (b) the model. The rat data show preference index for the correct responses, $(x - y)/(x + y)$, where x is the transitive response and y is the alternate, and the model results are in terms of proportion of correct responses. Results for three different training conditions in the model are shown: two sequential blocks of 50 trials each, 10 blocks of 10 trials, and fully interleaved. The intact model exhibits a relatively strong transitivity effect compared with the model with the hippocampal component removed (HL), and this effect is modulated by the recency of the trials containing the appropriate output responses as a function of the training conditions.

- Provide a theoretical framework that can accommodate the conflicting evidence on hippocampal conjunctive representations.
- Use this framework to identify better empirical tests of the conjunctive representations hypothesis (e.g., incidental conjunctive learning and contextual fear conditioning).
- Implement this framework in a computational neural network model that simulates a wide range of empirical data across different task paradigms while also making novel predictions.

We propose that the conflict between the conjunctive theory and the behavioral data can be resolved by developing a broader framework for understanding the division of labor between the cortex and the hippocampus. We adopt the general characterization of McClelland et al. (1995), where the cortex acquires information gradually to extract the generalities shared across different experiences, whereas the hippocampus acquires information rapidly and keeps specific events distinct. Our unique assumption is that both the cortex and hippocampus are able to store representations of stimulus conjunctions, but the cortex does so only when forced by the demands of the environment, such as in the case of nonlinear discrimination learning problems. In contrast, the hippocampus generally encodes stimulus conjunctions automatically as a by-product of the organism sampling its environment (but it also performs pattern completion when the inputs are sufficiently similar to stored representations).

Our computational models of both the cortex and hippocampus are based on a common set of principles embodied in the Leabra algorithm (O'Reilly, 1996b, 1998; O'Reilly & Munakata, 2000). These principles include the use of error-driven learning based on task demands, Hebbian learning that is sensitive to the co-occurrence of features, and inhibitory competition for producing sparse distributed representations. In this model, the hippocampus and cortex lie on a parametric continuum, with the hippocampus having both greater inhibitory competition and thus sparser representations, and a somewhat greater reliance on Hebbian as opposed to error-driven learning.

It is worth reiterating that, using one basic model, we were able to successfully simulate the results of experiments from a wide range of paradigms that have been used to evaluate the role of the hippocampus. These tasks range from complex nonlinear discrimination problems to the relatively simple paradigms of fear conditioning and habituation. We were also able to simulate the results of the complex transitive inference tasks that have been used to demonstrate memory flexibility. Indeed, our model suggests that basic pattern completion processes can provide the basis for the logical operations hypothesized to underlie transitive inference in animals. We now highlight some of the insights gained from this exercise and then consider a set of other important issues in subsequent sections, concluding with a discussion about other perspectives on the hippocampus.

Insights

Perhaps one of the most important insights gained from this exercise is the importance of differentiating between representations of stimulus conjunctions that are constructed in the service of solving discrimination problems (and thus influenced by error-driven learning pressures) and conjunctive representations that emerge automatically, rapidly, and incidentally from exposure to the environment. Failure to distinguish between these two cases has led to some of the past difficulties encountered in understanding the primary role of the hippocampus.

Although the contribution of the hippocampus in nonlinear discrimination problems is relatively small and the empirical data somewhat inconsistent, we nevertheless achieved useful insights into the critical features of different nonlinear discrimination problems that cause them to be more or less sensitive to hippocampal function. In addition to highlighting the importance of whether stimuli appear alone or in combination, we found that blocked versus interleaved training plays an important role in whether conjunctive representations are actually required, and thus whether the hippocampus makes an important contribution. We were able to make the novel prediction that hippocampal damage should not substantially impair learning of the blocked version of the NP problem.

In contextual fear conditioning, we verified a number of earlier suggestions about the role of the hippocampus in constructing a unitary representation of context. Some of these suggestions (e.g., Rudy & O'Reilly, 1999) were based on our theoretical framework and constitute important insights into both the conjunctive nature of the hippocampal context representations and the role of pattern completion in producing generalized fear conditioning.

We found that the purported importance of the hippocampus in enabling flexible behavior (e.g., Eichenbaum, 1992; O'Keefe & Nadel, 1978) appears to be explainable in terms of the pattern completion abilities of the hippocampus. Specifically, we showed that the transitivity tests performed on rats by Bunsey and Eichenbaum (1996) and Dusek and Eichenbaum (1997) could be simulated by hippocampal pattern completion in our model. We achieved several important insights into the influence of the training procedures on producing the "flexible" behavior and generated several novel predictions regarding the effects of manipulations of these procedures.

We suggest that it may be more productive to focus on the more mechanistic principle of pattern completion instead of the more abstract notion of flexibility in conceptualizing the unique behavioral contributions of the hippocampus. Furthermore, we also note that models of slow, integrative cortical learning are capable of demonstrating flexibility in the form of generalizing to novel inputs (e.g., pronouncing novel nonwords; Plaut, McClelland, Seidenberg, & Patterson, 1996). Indeed, one of the primary advantages of this slow, integrative learning is that it facilitates generalization based on the regularities extracted from a large number of prior experiences. Thus, the overall behavioral flexibility of an organism can presumably be subserved by multiple underlying mechanisms, each with different properties.

Human Hippocampal Function

We have focused the present applications of the model on the animal literature because it provides fertile ground for testing mechanistic theories of hippocampal function, but we believe that our general framework also will be useful for understanding the nature of human memory. Consistent with this view, we note that Squire (1992) has suggested that the conjunctive learning mechanism supported by the hippocampus underlies human declarative memory. The notion of a conjunctive binding mechanism is also implicit in Tulving's (1972) model of human episodic memory (see Mishkin et al., 1998). Moreover, it is generally appreciated that the basic anatomy of the hippocampus is preserved across

rodents as well as primates, including humans, so aside from differences in overall numbers of neurons and perhaps some scaling of different areas, the human hippocampal circuit appears to be consistent with the basic principles of our framework. Thus, we are optimistic that the general principles captured in our model can be successfully applied to a range of different human memory phenomena; such efforts are underway (Norman, O'Reilly, & Huber, 2000; O'Reilly et al., 1998).

Cortical Contributions to Memory Phenomena

The assumption that the cortex learns gradually is central to our model. However, there are preserved memory functions in human amnesics such as the single trial priming effect (Graf, Squire, & Mandler, 1984; Schacter & Graf, 1986) that appear to violate our key assumption about cortical learning. We suggest that such effects reflect the impact that small incremental changes can have on existing representations. In support of this position, several different neural network models have shown that slow learning rates can exhibit measurable effects on existing representations. Such effects result from slightly facilitating the processing of a stimulus or by shifting the balance of strength among a set of existing representations (e.g., Becker, Moscovitch, Behrmann, & Joordens, 1997; McClelland & Rumelhart, 1986; O'Reilly & Munakata, 2000).

Furthermore, we have recently shown that these same small effects can result in good recognition memory performance (Norman et al., 2000), which can account for findings of relatively preserved recognition memory with selective hippocampal damage (Aggleton & Brown, 1999; Aggleton & Shaw, 1996; Holdstock et al., in press; Murray & Mishkin, 1986; Squire & Zola-Morgan, 1991; Vargha-Khadem et al., 1997; Zola-Morgan, Squire, Amaral, & Suzuki, 1989). Our current work shows many important differences between the cortical and hippocampal contributions to recognition memory (Norman et al., 2000; O'Reilly et al., 1998), suggesting that this domain will be particularly informative for further tests of our general framework.

Finally, other examples of rapid learning, such as taste aversion learning or fear conditioning, also appear to violate the slow cortical learning hypothesis. However, the rapid learning seen in these domains is generally believed to be the product of specialized evolutionary adaptations (Bolles, 1970; Garcia, McGowan, & Green, 1972; Seligman, 1970).

Other Perspectives on the Hippocampus

We have discussed a number of different perspectives on the hippocampus and showed how many of them are generally consistent with our framework. Nevertheless, some important similarities and differences should be highlighted (also see McClelland et al., 1995, for other relevant comparisons).

Similarities. Our theoretical framework for understanding the division of labor between the cortex and hippocampus is remarkably similar to that developed by O'Keefe and Nadel (1978) to differentiate their local and taxon systems. We already noted that we view their idea that the hippocampal-dependent locale system supports the acquisition and memory of maplike representations as being related to our idea that the hippocampus is important for learning stimulus conjunctions. Although much of the subsequent discussion of O'Keefe and Nadel's ideas in the literature has focused on this spatial representation idea, they also made distinctions between the taxon and locale system along other important dimensions:

> *Learning rate.* The locale system is viewed as rapidly storing new information, whereas the taxon system learns and unlearns by slow increments.
>
> *Motivation.* The two systems operate under different motivational conditions. The locale system is fundamentally connected to exploration and much of what it encodes occurs as a result of novelty directed behavior. Taxon learning, however, is motivated to learn in the service of problem solving or achieving goals and is therefore sensitive to the reinforcement contingencies associated with behavior.
>
> *Susceptibility to interference.* The two systems are differentially susceptible to associative interference. The locale system is suited to reduce interference because it encodes experiences in unitary maplike formats that emphasize the uniqueness of the episode, preventing interference from other similar experiences.

Each of these dimensions apply to our distinctions between the cortical and hippocampal systems: (a) The cortical system learns slowly compared with the hippocampal system; (b) the hippocampus is biased to automatically form conjunctive representations, whereas the cortex must generally be forced by task demands to develop such representations; and (c) the hippocampus uses pattern separation to enable rapid learning of arbitrary information without suffering from undue interference. Thus, although we developed our framework largely from computational principles, we have arrived at similar conclusions. To the degree that our model has captured many important findings in the modern literature, O'Keefe and Nadel (1978) clearly anticipated the critical features of a successful mechanistic model. Nevertheless, we differ importantly from O'Keefe and Nadel because they restricted the content encoded by the hippocampus to spatial information whereas our view is more inclusive, allowing for the storage of nonspatial and spatial conjunctions.

We also noted that the ideas of Sherry and Schacter (1987) are very similar to the complementary learning systems framework of McClelland et al. (1995). Again, this demonstrates that our computational principles have converged on ideas that can also be motivated by other considerations. Also, a number of computational models of hippocampal function have embraced some of the assumptions that are central to our models (e.g., Alvarez & Squire, 1994; Burgess & O'Keefe, 1996; Hasselmo, 1996; Moll & Miikkulainen, 1997; Touretzky & Redish, 1996; Treves & Rolls, 1994; Wu, Baxter, & Levy, 1996).

Differences. Perhaps the clearest contrast between our perspective and some others centers on the learning capacities of the cortex. Several mechanistic accounts of the hippocampus assume that the cortex is a repository for knowledge and must rely on other brain structures such as the hippocampus to acquire its impressive cognitive functions (Gluck & Myers, 1993; Rolls, 1990; Schmajuk & DiCarlo, 1992; Wickelgren, 1979). For example, Gluck and Myers (1993) assumed that the hippocampus uses a relatively powerful learning mechanism (error backpropogation) and that the cortex is effectively a slave to this hippocampal mechanism for anything but the most simple forms of learning. In a related view, Schmajuk and DiCarlo (1992) assumed that the hippocampus

plays an essential role in enabling error-driven modifications of cortical representations to occur.

This codependent view of cortical learning, however, does not appear to be tenable. There is impressive evidence of sophisticated learning by amnesic humans (e.g., Knowlton, Squire, & Gluck, 1994; Squire & Knowlton, 1995; Vargha-Khadem et al., 1997). In addition, the animal literature reviewed here and elsewhere (Rudy & Sutherland, 1995) indicates that the cortex does not depend on the hippocampus to solve many complex nonlinear discrimination tasks. A strength of our model is that it assumes that the cortex, without the hippocampus, is capable of quite sophisticated learning. Thus, it can account for the fact that animals with damage to the hippocampus can solve complex nonlinear discriminations and is positioned to explain other complex learning phenomena displayed by patients with damage to the hippocampus.

Conclusion

The idea that the hippocampus encodes representations of stimulus conjunctions is common to many theories. Our analysis of the literature, however, indicated that, unconstrained, this idea cannot be correct. To resolve the tension created by this analysis, we placed this idea into a broader framework that addressed fundamental differences in cortical and hippocampal learning systems. This framework recognizes that both systems can support the learning of stimulus conjunctions but that the hippocampus does so rapidly and automatically simply as a consequence of the organism exploring and attending to its environment, whereas the cortex does so gradually when driven by the demands of the task. We embedded these ideas into a biologically based computational model. This model was able to simulate a wide range of findings and appears to resolve the problems created by the finding that the hippocampus is not necessary to solve problems that require conjunctive representations. Nevertheless, much work needs to be done to fully explore the ideas laid out in this article. We hope that this first step provides a solid foundation for future research.

References

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

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

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.

Alvarez, P., & Squire, L. R. (1994). Memory consolidation and the medial temporal lobe: A simple network model. *Proceedings of the National Academy of Sciences, USA, 91,* 7041–7045.

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.

Becker, S., Moscovitch, M., Behrmann, M., & Joordens, S. (1997). Long-term semantic priming: A computational account and empirical evidence. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23,* 1059–1082.

Bolles, R. C. (1970). Species specific defense reactions and avoidance learning. *Psychological Review, 77,* 32–48.

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,* 255–257.

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

Bussey, T. J., Warburton, E. C., Aggleton, J. P., & Muir, J. L. (1999). Fornix lesions can facilitate acquisition of the transverse patterning task: A challenge for "configural" theories of hippocampal function. *Journal of Neuroscience, 18,* 1622–1631.

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

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

Cohen, 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* (pp. 267–296). Mahwah, New Jersey: Erlbaum.

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.

Davis, M. (1992). The role of the amygdala in conditioned fear. In J. P. Aggleton (Ed.), *The amygdala: Neurobiological aspects of emotion, memory, and mental dysfunction* (pp. 255–305). New York: Wiley-Liss.

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,* 762–771.

Eichenbaum, H. (1992). The hippocampal system and declarative memory in animals. *Journal of Cognitive Neuroscience, 4,* 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). New York: Wiley.

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

Fiez, J. A. (1996). Cerebellar contributions to cognition. *Neuron, 16,* 13–15.

Gaffan, D. (1974). Recognition impaired and association intact in the memory of monkeys after transection of the fornix. *Journal of Comparative and Physiological Psychology, 86,* 1100–1109.

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

Gao, J. H., Parsons, L. M., & Fox, P. T. (1996, April 26). Cerebellum implicated in sensory acquisition and discrimination rather than motor control. *Science, 272,* 545–547.

Garcia, J., McGowan, B. K., & Green, K. F. (1972). Biological constraints on conditioning. In M. E. P. Seligman & J. L. Hager (Eds.), *Biological boundaries of learning* (pp. 21–43). New York: Appleton-Century-Crofts.

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,* 1171–1183.

Graf, P., Squire, L. R., & Mandler, G. (1984). The information that amnesic patients do not forget. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 164–178.

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

Hasselmo, M. E. (1996). Neuromodulation and cortical function: Modeling the physiological basis of behavior. *Behavioural Brain Research, 67,* 1–27.

Hasselmo, M. E., & Wyble, B. (1997). Free recall and recognition in a network model of the hippocampus: Simulating effects of scopolamine on human memory function. *Behavioural Brain Research, 89,* 1–345.

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.

Holdstock, J. S., Mayes, A. R., Roberts, N., Cezayirli, E., Isaac, C. L., O'Reilly, R. C., & Norman, K. A. (in press). Memory dissociations following human hippocampal damage. *Hippocampus.*

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

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

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.

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, May). Modality-specific retrograde amnesia of fear. *Science, 256,* 675–677.

Knowlton, B. J., Squire, L. R., & Gluck, M. A. (1994). Probabilistic category learning in amnesia. *Learning and Memory, 1,* 1–15.

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

LeDoux, J. E. (1992). Brain mechanisms of emotion and emotional learning. *Current Opinion in Neurobiology, 2,* 191–197.

Levy, W. B. (1989). A computational approach to hippocampal function. In R. D. Hawkins & G. H. Bower (Eds.), *Computational models of learning in simple neural systems* (pp. 243–304). San Diego, CA: Academic Press.

Lubow, R. E. (1989). *Latent inhibition and conditioned attention theory.* Cambridge, England: 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. (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 connectionist models of learning and memory. *Psychological Review, 102,* 419–457.

McClelland, J. L., & Rumelhart, D. E. (1986). A distributed model of human learning and memory. In J. L. McClelland, D. E. Rumelhart, & PDP Research Group (Eds.), *Parallel distributed processing: Vol. 2. Psychological and biological models* (pp. 170–215). Cambridge, MA: MIT Press.

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.

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

McNaughton, B. L., & Morris, R. G. M. (1987). Hippocampal synaptic

enhancement and information storage within a distributed memory system. *Trends in Neurosciences, 10,* 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* (pp. 1–63). Hillsdale, NJ: Erlbaum.

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.

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

Mishkin, M., Malamut, B., & Bachevalier, J. (1984). Memories and habits: Two neural systems. In G. Lynch, J. L. McGaugh, & N. M. Weinberger (Eds.), *Neurobiology of learning and memory* (pp. 65–77). New York: Guilford.

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: Guilford.

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

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

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 preservation and implications for object permanence theories: A PDP model of the $A\bar{B}$ 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.

Murray, E. A., & Mishkin, M. (1986). Visual recognition in monkeys following rhinal cortical ablations combined with either amydalectomy or hippocampectomy. *Journal of Neuroscience, 6,* 1991–2003.

Nadel, L. (1994). Multiple memory systems: What and why, and update. In D. L. Schacter & E. Tulving (Eds.), *Memory systems 1994* (pp. 39–63). Cambridge, MA: MIT Press.

Nadel, L., & O'Keefe, J. (1974). The hippocampus in pieces and patches: An essay on modes of explanation in physiological psychology. In R. Bellaires & E. G. Gray (Eds.), *Essays on the nervous system: A festschrift for Professor J. Z. Young* (pp. 367–390). Oxford, England: Clarendon Press.

Norman, K. A., O'Reilly, R. C., & Huber, D. E. (2000, March). *Modeling neocortical contributions to recognition memory.* Paper presented at the annual meeting of the Cognitive Neuroscience Society, San Francisco, CA.

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

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 cognitive map.* Oxford, England: Oxford University Press.

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

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

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

O'Reilly, R. C. (in press). Generalization in interactive networks: The benefits of inhibitory competition and Hebbian learning. *Neural Computation.*

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

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain.* Cambridge, MA: MIT Press.

O'Reilly, R. C., Norman, K. A., & McClelland, J. L. (1998). A hippocampal model of recognition memory. In M. I. Jordan, M. J. Kearns, & S. A. Solla (Eds.), *Advances in neural information processing systems 10* (pp. 73–79). Cambridge, MA: MIT Press.

Packard, M. G., Hirsh, R., & White, N. M. (1989). Differential effects of fornix and caudate nucleus lesions on two radial maze tasks: Evidence for multiple memory systems. *Journal of Neuroscience, 9,* 1465–1472.

Penfield, W., & Milner, B. (1958). Memory deficits produced by bilateral lesions in the hippocampal zone. *Archives of Neurology and 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.

Reed, J. M., & Squire, L. R. (1999). Impaired transverse patterning to human amnesia is a special case of impaired memory for two-choice discrimination tasks. *Behavioral Neuroscience, 113,* 3–9.

Risold, P. Y., & Swanson, L. W. (1996, June 7). 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* (pp. 73–90). 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. (1999). Contextual fear conditioning, conjunctive representations, pattern completion, and the hippocampus. *Behavioral Neuroscience, 113,* 867–880.

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. (1986). Learning

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

Rumelhart, D. E., McClelland, J. L., & PDP Research Group (Eds.). (1986). *Parallel distributed processing: Vol. 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: Vol. 1. Foundations* (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,* 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.

Seligman, M. E. P. (1970). On the generality of the laws of learning. *Psychological Review, 77,* 406–418.

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

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

Squire, L. R. (1987). *Memory and brain.* Oxford, England: 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., & Knowlton, B. J. (1995). Learning about categories in the absence of memory. *Proceedings of the National Academy of Sciences, 92,* 12470–12474.

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

Squire, L. R., Zola-Morgan, S., & Chen, K. S. (1988). Human amnesia and animal models of amnesia: Performance of amnesic patients on tests designed for the monkey. *Behavioral Neuroscience, 102,* 210–221.

Squire, L. R., & Zola-Morgan, S. M. (1991, September). The medial temporal lobe memory system. *Science, 253,* 1380–1386.

Sutherland, R. J., McDonald, R. J., Hill, C. R., & Rudy, J. W. (1989). 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,* 129–144.

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

Suzuki, W. A. (1996). The anatomy, physiology and functions of the perirhinal cortex. *Current Opinion in Neurobiology, 6,* 179–186.

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

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

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

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

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

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

Tulving, E., & Markowitsch, H. J. (1998). Episodic and declarative memory: Role of the hippocampus. *Hippocampus, 8,* 198–204.

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, July). Differential effects of early hippocampal pathology on episodic and semantic memory. *Science, 277,* 376–380.

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.

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

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

Wood, E. R., Dudchenko, P. A., & Eichenbaum, H. (1999). The global record of memory in hippocampal neuronal activity. *Nature, 397,* 613–616.

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

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

Zola-Morgan, S., Squire, L. R., Amaral, D. G., & Suzuki, W. A. (1989). Lesions of perirhinal and parahippocampal cortex that spare the amygdala and hippocampal formation produce severe memory impairment. *Journal of Neuroscience, 9,* 4355–4370.

(Appendix follows)

Appendix

Computational Mechanisms

This appendix describes the computational details of the Leabra algorithm that was used in the simulations.

Pseudocode

The pseudocode for Leabra is given here, showing exactly how the pieces of the algorithm described in more detail in the subsequent sections fit together.

Outer loop: Iterate over events (trials) within an epoch. For each event:

1. Iterate over minus and plus phases of settling for each event.
   (a) At start of settling, for all units:
      i. Initialize all state variables (activation, v_m, etc).
      ii. Apply external patterns (clamp input in minus, input & output in plus).
   (b) During each cycle of settling, for all nonclamped units:
      i. Compute excitatory netinput, $g_e(t)$ or η_j (Equation A3).
      ii. Compute kWTA inhibition for each layer, based on g_i^Θ (Equation A6):
         A. Sort units into two groups based on g_i^Θ: top k and remaining $k + 1$ to n.
         B. Set inhib conductance g_i between g_k^Θ and g_{k+1}^Θ (Equation A5).
      iii. Compute point-neuron activation combining excitatory input and inhibition (Equation A1).
   (c) After settling, for all units: Record final settling activations as either minus or plus phase (y_j^- or y_j^+).
2. After both phases update the weights (based on linear current weight values), for all connections:
   (a) Compute error-driven weight changes (Equation A7) with soft weight bounding (Equation A9).
   (b) Compute Hebbian weight changes from plus-phase activations (Equation A8).
   (c) Compute net weight change as weighted sum of error-driven and Hebbian changes (Equation A10).
   (d) Increment the weights according to net weight change, and apply contrast-enhancement (Equation A11).

Point-Neuron Activation Function

Leabra uses a *point-neuron* activation function that models the electrophysiological properties of real neurons while simplifying their geometry to a single point. This function is nearly as simple computationally as the standard sigmoidal activation function, but the more biologically based implementation makes it considerably easier to model inhibitory competition, as described. Further, using this function enables cognitive models to be more easily related to more physiologically detailed simulations, thereby facilitating bridge building between biology and cognition.

The membrane potential V_m is updated as a function of ionic conductances g with reversal (driving) potentials E as follows:

$$\frac{dV_m(t)}{dt} = \tau \sum_c g_c(t)\bar{g}_c[E_c - V_m(t)] \tag{A1}$$

with three channels (c) corresponding to e excitatory input, l leak current, and i inhibitory input. Following electrophysiological convention, the overall conductance is decomposed into a time-varying component $g_c(t)$ computed as a function of the dynamic state of the network and a constant $\bar{g}_c$ that controls the relative influence of the different conductances. The equilibrium potential can be written in a simplified form by setting the excitatory driving potential (E_e) to 1 and the leak and inhibitory driving potentials (E_l and E_i) of 0:

$$V_m^\infty = \frac{g_e\bar{g}_e}{g_e\bar{g}_e + g_l\bar{g}_l + g_i\bar{g}_i}, \tag{A2}$$

which shows that the neuron is computing a balance between excitation and the opposing forces of leak and inhibition. This equilibrium form of the equation can be understood in terms of a Bayesian decision-making framework (O'Reilly & Munakata, 2000).

The excitatory net input/conductance $g_e(t)$ or η_j is computed as the proportion of open excitatory channels as a function of sending activations times the weight values:

$$\eta_j = g_e(t) = \langle x_i w_{ij} \rangle = \frac{1}{n}\sum_i x_i w_{ij}. \tag{A3}$$

The inhibitory conductance is computed via the kWTA function described in the next section, and leak is a constant.

Activation communicated to other cells (y_j) is a thresholded (Θ) sigmoidal function of the membrane potential with gain parameter γ:

$$y_j(t) = \frac{1}{\left(1 + \frac{1}{\gamma[V_m(t) - \Theta]_+}\right)}, \tag{A4}$$

where $[x]_+$ is a threshold function that returns 0 if $x < 0$ and x if $X > 0$. This sharply thresholded function is convolved with a Gaussian noise kernel ($\sigma = .005$), which reflects the intrinsic processing noise of biological neurons. This produces a less discontinuous deterministic function with a softer threshold that is better suited for graded learning mechanisms (e.g., gradient descent).

kWTA Inhibition

Leabra uses a kWTA function to achieve sparse distributed representations. Although two different versions are possible (see O'Reilly & Munakata, 2000, for details), only the simpler, more rigid form was used in the present simulations. A uniform level of inhibitory current for all units in the layer is computed as follows:

$$g_i = g_{k+1}^\Theta + q(g_k^\Theta - g_{k+1}^\Theta), \tag{A5}$$

where $0 < q < 1$ is a parameter for setting the inhibition between the upper bound of g_k^Θ and the lower bound of g_{k+1}^Θ. These boundary inhibition values are computed as a function of the level of inhibition necessary to keep a unit right at threshold:

$$g_i^\Theta = \frac{g_e^*\bar{g}_e(E_e - \Theta) + g_l\bar{g}_l(E_l - \Theta)}{\Theta - E_i}, \tag{A6}$$

where g_e^* is the excitatory net input without the bias weight contribution—this allows the bias weights to override the kWTA constraint.

In the basic version of the kWTA function used here, which is relatively rigid about the kWTA constraint, g_k^Θ and g_{k+1}^Θ are set to the threshold inhibition value for the kth and $k + 1$th most excited units, respectively. Thus, the inhibition is placed exactly to allow k units to be above threshold and the remainder below threshold. For this version, the q parameter is almost always .25, allowing the kth unit to be sufficiently above the inhibitory threshold.

Activation dynamics similar to those produced by the kWTA function have been shown to result from simulated inhibitory interneurons that

project both feedforward and feedback inhibition (O'Reilly & Munakata, 2000). Thus, although the kWTA function is somewhat biologically implausible in its implementation (e.g., requiring global information about activation states and using sorting mechanisms), it provides a computationally effective approximation to biologically plausible inhibitory dynamics.

Error-Driven Learning

Leabra uses the symmetric midpoint version of the GeneRec algorithm (O'Reilly, 1996a), which is functionally equivalent to the deterministic Boltzmann machine and contrastive Hebbian learning (Hinton, 1989; Movellan, 1990). The network settles in two phases, an expectation (minus) phase, where the network's actual output is produced, and an outcome (plus) phase, where the target output is experienced, and then computes a simple difference of a pre- and postsynaptic activation product across these two phases:

$$\Delta_{\text{err}} w_{ij} = (x_i^+ y_j^-) - (x_i^- y_j^-) \tag{A7}$$

for sending unit x_i and receiving unit y_j in the two phases.

Hebbian Learning

The simplest form of Hebbian learning adjusts the weights in proportion to the product of the sending (x_i) and receiving (y_j) unit activations: $\Delta w_{ij} = x_i y_j$. The weight vector is dominated by the principal eigenvector of the pairwise correlation matrix of the input, but it also grows without bound. Leabra uses essentially the same learning rule used in competitive learning or mixtures-of-Gaussians (Nowlan, 1990; Rumelhart & Zipser, 1986), which can be seen as a variant of the Oja normalization (Oja, 1982):

$$\Delta_{\text{hebb}} w_{ij} = x_i^+ y_j^+ - y_j^+ w_{ij} = y_j^+ (x_i^+ - w_{ij}). \tag{A8}$$

Rumelhart and Zipser (1986) and O'Reilly and Munakata (2000) showed that, when activations are interpreted as probabilities, this equation converges on the conditional probability that the sender is active given that the receiver is active.

Combining Error-Driven and Hebbian Learning

Error-driven and Hebbian learning are combined additively at each connection to produce a net weight change. Two equations are needed, a soft weight bounding equation to keep the error-driven component within the same 0–1 range of the Hebbian term and the combination equation.

Soft weight bounding with exponential approach to the 0–1 extremes is implemented using

$$\Delta_{\text{sberr}} w_{ik} = [\Delta_{\text{err}}]_+ (1 - w_{ik}) + [\Delta_{\text{err}}]_- w_{ik}, \tag{A9}$$

where Δ_{err} is the error-driven weight change, Δ_{sberr} is the soft-bounded weight change, and the $[x]_+$ operator returns x if $x > 0$ and 0 otherwise, while $[x]_-$ does the opposite, returning x if $x < 0$ and 0 otherwise.

The net weight change equation combining error-driven and Hebbian learning (which also includes the learning rate parameter ϵ) uses a normalized mixing constant k_{hebb}:

$$\Delta w_{ij} = \epsilon[k_{\text{hebb}}(\Delta_{\text{hebb}}) + (1 - k_{\text{hebb}})(\Delta_{\text{sberr}})]. \tag{A10}$$

To increase the influence of Hebbian learning in the hippocampus relative to the cortex, k_{hebb} for the hippocampus was .05, while it was .02 for the cortex.

Weight Contrast Enhancement

One limitation of the Hebbian learning algorithm is that the weights linearly reflect the strength of the conditional probability. This linearity can limit the network's ability to focus on only the strongest correlations while ignoring weaker ones. To remedy this limitation, we introduce a contrast enhancement function that magnifies the stronger weights and shrinks the smaller ones in a parametric, continuous fashion. This contrast enhancement is achieved by passing the linear weight values computed by the learning rule through a sigmoidal nonlinearity of the following form:

$$\hat{w}_{ij} = \frac{1}{1 + \left(\theta \frac{w_{ij}}{1 - w_{ij}}\right)^{-\gamma}}, \tag{A11}$$

where $\hat{w}_{ij}$ is the contrast-enhanced weight value, and the sigmoidal function is parameterized by an offset θ and a gain γ (standard defaults of 1.25 and 6, respectively, used here).

Received January 26, 1999
Revision received June 30, 2000
Accepted July 8, 2000 ■