Psychological Review, in press

Why there are Complementary Learning Systems in the Hippocampus and Neocortex: Insights from the Successes and Failures of Connectionist Models of Learning and Memory

James L. McClelland
Carnegie Mellon University and the
Center for the Neural Basis of Cognition

Bruce L. McNaughton
University of Arizona

Randall C. O'Reilly
Carnegie Mellon University and the
Center for the Neural Basis of Cognition

Damage to the hippocampal system disrupts recent memory but leaves remote memory intact. Our account of this suggests that memories are first stored via synaptic changes in the hippocampal system; that these changes support reinstatement of recent memories in the neocortex; that neocortical synapses change a little on each reinstatement; and that remote memory is based on accumulated neocortical changes. Models that learn via adaptive changes to connections help explain this organization. These models discover the structure in ensembles of items if learning of each item is gradual and interleaved with learning about other items. This suggests that neocortex learns slowly to discover the structure in ensembles of experiences. The hippocampal system permits rapid learning of new items without disrupting this structure, and reinstatement of new memories interleaves them with others to integrate them into structured neocortical memory systems.

One of the most striking neuropsychological phenomena ever reported is the dramatic amnesia produced by bilateral lesions to the hippocampus and related temporal lobe structures (Scoville & Milner, 1957). A crucial aspect of this phenomenon is temporally graded retrograde amnesia. Considerable evidence now supports the conclusion that the influence of the hippocampal system on the ability to exploit information derived from past experience in a wide range of tasks is temporally circumscribed: Performance is impaired if the hippocampal system is damaged before or within a window of time after the initial experience; but if the hippocampal system is left intact both during the experience and for a period of time thereafter, subsequent damage may have little or no impact on performance.

This change in dependence on the hippocampal system over time appears to be a slow, gradual process. This gradual change has often been called consolidation, but the term really only labels the phenomenon. In this paper, we focus on consolidation, and consider what produces it and why it occurs. We ask: Is the phenomenon a reflection of an arbitrary property of the nervous system, or does it reflect some crucial aspect of the mechanisms of learning and memory? Is the fact that consolidation can take quite a long time—up to 15 years or more in some cases—just an arbitrary parameter, or does it reflect an important design principle?

We begin with a brief overview of the neuropsychology of memory, emphasizing the temporally circumscribed role of the hippocampal system, and elaborate one possible account of the functional organization of memory that is broadly consistent with the neuropsychological evidence, as well as aspects of the underlying anatomy and physiology. We then describe results from connectionist modeling research that suggest reasons for this organization and for the phenomenon of gradual consolidation. From the insights gained through the consideration of these models we develop illustrative simulation models of the phenomenon of temporally graded retrograde amnesia. These

The work reported herein was supported by Grants MH00385 and MH47566 to the first author, by Grant MH46823 to the second author, by a collaborative grant from the Human Frontiers Program, and by a Center Grant from the McDonnell-Pew Program in Cognitive Neuroscience to the University of Arizona. The third author was partially supported by an ONR Fellowship. We thank Carol Barnes, Sue Becker, Marlene Behrmann, Loren Frank, Mark Gluck, Robert French, Patrick Lynn, Mort Mishkin, Catherine Myers, David Plaut, William Skaggs, Craig Stark, Robert Sutherland, Larry Squire, and Matthew Wilson for useful discussions and/or comments on earlier drafts of this article. Lynn Nadel's contributions to the initial formulation of some of these ideas and to several revisions of the manuscript were extensive and are gratefully acknowledged. The computational arguments developed here were sketched in an Society for Neurosciences abstract (McClelland, McNaughton, O'Reilly & Nadel, 1992) and an invited review of several projects involving the first author that referred to the present article as work in progress (McClelland, 1994). Preliminary versions of the simulations of retrograde amnesia in rats and monkeys were previously presented in a Psychonomics abstract (McClelland, McNaughton & O'Reilly, 1993).

Correspondence concerning this manuscript may be sent to the first author at the Department of Psychology, Carnegie Mellon University, Pittsburgh, PA, 15213 or via email to mcclelland+@cmu.edu.

are not detailed neural models; rather they illustrate at an abstract level what we take consolidation to be about. We discuss the implications of our view of the role of consolidation for findings related to age, species, and task differences in neocortical learning and for the form of representations used in the hippocampus, and we conclude with a comparison of our views to those of others who have theorized about the role of the hippocampal system in learning and memory. While there are many points of compatibility, our approach differs from some others in treating gradual consolidation as reflecting a principled aspect of the design of the mammalian memory system.

Role of the Hippocampal System in Learning and Memory

The phrase *the hippocampal system* is widely used to refer to a system of interrelated brain regions found in a range of mammalian species that appear to play a special role in learning and memory. The exact boundaries of the hippocampal system are difficult to define, but it includes at least the hippocampus itself—the CA1-3 fields of Ammon's Horn and the dentate gyrus—the subicular complex and the entorhinal cortex. It probably also encompasses adjacent structures including the perirhinal and parahippocampal cortices.

The literature on the effects of damage to the hippocampal system is quite vast. Here we summarize what we believe are the main points.

(1) An extensive lesion of the hippocampal system can produce a profound deficit in new learning, while leaving other cognitive functions and memory performance based on material acquired well before the lesion apparently normal. Dramatic evidence of this was first reported by Scoville and Milner (1957) in their description of the anterograde amnesia produced in patient HM due to bilateral removal of large portions of the hippocampal system and other temporal lobe structures. HM presented initially with a profound deficit in memory for events that occurred either after the lesion or during the weeks and months prior to it, with intact intellectual function and information processing skills and apparent sparing of his memory for more remote time periods.

(2) The effects of lesions to the hippocampal system appear to be selective to certain forms of learning. In humans, the hippocampal system appears to be essential for the rapid formation of comprehensive associations among the various elements of specific events and experiences, in a form sufficient to sustain an explicit (Schacter, 1987) retrieval of the contents of the experience, so that they can be attested (explicitly recognized as memories), verbally described, or flexibly used to govern subsequent behavior. Cohen and Squire (1980) introduced the term *declarative* memory to encompass these forms of memory. Included in the category of declarative memories are *episodic memories* (Tulving, 1983)—memories for the specific contents of individual episodes or events—as well as what are generally termed semantic memories, including knowledge of the meanings of words, factual information, and encyclopedic memories (see Squire, 1992, for a recent discussion). A paradigm example of this form of memory is paired-associate learning of arbitrary word pairs. Prior associations to the cue word are unhelpful in this task, which depends on recall of the word that previously occurred with the cue word in the list study context. Hippocampal system lesions produce profound impairments in learning arbitrary paired associates (Scoville & Milner, 1957). However, it should be noted that deficits are not apparently restricted to tasks that rely on memories that are explicitly accessed and used to govern task performance. For example, amnesics are also impaired in acquisition of arbitrary new factual information, whether or not the use of this information is accompanied by deliberate or conscious recollection of previous experience (Shimamura & Squire, 1987). Also, normal subjects show sensitivity to novel associations after a single presentation in stem completion tasks, but profound amnesics do not (Schacter & Graf, 1986; Shimamura & Squire, 1989). While recent evidence (Bowers & Schacter, 1993) suggests that normal subjects who show sensitivity to novel associations are conscious of having accessed these associations on at least some trials, sensitivity to novel associations is dissociable in several ways from standard measures of explicit or declarative memory (Graf & Schacter, 1987; Schacter & Graf, 1989). Thus, at this point the extent to which deficits in amnesics are restricted to tasks that depend on conscious access to the contents of prior episodes or events is unclear. It does appear that in humans an intact hippocampal system is necessary for the formation of an association between arbitrarily-paired words that is sufficiently strong after a single presentation to have any effect on subsequent performance, whether explicit memory is involved or not.

In the animal literature, the exact characterization of the forms of learning that depend on the hippocampal system remains a matter of intense investigation and debate. Sutherland and Rudy (1989) suggest that the hippocampal system is crucial for learning to make appropriate responses that depend not on individual cues but on specific combinations or conjunctions of cues—what they call *cue configurations*. The paradigm example of a task depending on cue configurations is the negative patterning task, in which animals receive reward for operant responses to a light and a tone but not the tone-light compound. Hippocampal system lesions lead to deficits in responding differently to the compound than to the individual cues (Rudy & Sutherland, 1989). Cohen and Eichenbaum (1993) have emphasized the importance of the hippocampal system for flexible access to memory traces, a characteristic that may be closely related to declarative memory in humans. A major alternative viewpoint is that of O'Keefe and Nadel (1978), who have suggested that the hippocampal system is especially relevant in the formation of memories involving places or locations in the environment, and there is a vast body of evidence that spatial learning is impaired following hippocampal lesions in rats. One view of spatial learning that is compatible with both the Sutherland and Rudy (1989) and the O'Keefe and Nadel (1978) theories is that place learning involves forming configural associations of locations and movements that would enable prediction of the spatial consequence of a given movement in a given spatial context (McNaughton, Leonard, & Chen, 1989). This can be seen as a special case of the Sutherland and Rudy (1989) theory, and it is possible that it may be the evolutionary forerunner of the more general processing capability. In any case, increasing evidence suggests that damage restricted to the hippocampus impacts on tasks that require the animal to learn responses specific to particular nonspatial combinations of cues, or to specific contexts, as well as tasks that depend on learning to navigate in a previously unfamiliar spatial environment (Jarrard, 1993; Rudy & Sutherland, 1994).[1] More extensive lesions of the hippocampal system lead to deficits in a broader range of tasks. In some cases, selective lesions to just the hippocampus produce little or no effect, though performance is severely disturbed by a complete lesion of the entire hippocampal system (see Eichenbaum, Otto, & Cohen, 1994, and Jarrard, 1993, for reviews).

(3) Some kinds of learning appear to be completely unaffected by hippocampal system lesions. Squire (1992) characterizes these forms of memory as *non-declarative* or *implicit* (Schacter, 1987), emphasizing that they influence behavior without depending on conscious or deliberate access to memory for the contents of the events that led to these influences. Another characterization emphasizes inflexibility of use of such memories; they appear to influence behavior maximally when there is a close match between the processing carried out during the learning event and the processing carried out when the later influence of the learning event is assessed (Cohen & Eichenbaum, 1993). This greater specificity appears to characterize implicit memory as it is observed in normals as well as amnesics (Schacter, 1987). Examples of forms of learning that are spared are gradually acquired skills that emerge over several sessions of practice, such as the skill of tracing a figure viewed in a mirror (Milner, 1966), reading mirror-reversed print (Cohen & Squire, 1980), or anticipating subsequent items in a sequence governed by a complex stochastic grammar (Cleeremans, 1993). Hippocampal patients also appear to be spared in their ability to learn the structure common to a set of items: They are as good as normals in judging whether particular test items come from the same prototype, or were generated by the same finite-state grammar, as the members of a previously studied list (Knowlton, Ramus, & Squire, 1992; Knowlton & Squire, 1993). Spared learning is also exhibited in repetition priming tasks. These are tasks that require subjects to emit some response already within their capabilities, such as naming a word or picture (Milner, Corkin, & Teuber, 1968), reading aloud a pronounceable nonword (Haist, Musen, & Squire, 1991), or completing a word fragment with a lexically valid completion (Graf, Squire, & Mandler, 1984). Repetition priming is exhibited when the subject is later required to process a previously presented item, and a single prior presentation is often sufficient. In many such tasks, hippocampal patients appear indistinguishable from normals in the extent to which they show facilitation from prior presentations, as long as care is taken to avoid the possibility that explicit recall is used to aid performance. Hippocampal patients exhibit spared priming of existing associations (i.e., and increase in the likelihood of producing *table* when giving a free-associate to *chair* after prior presentation of *table* and chair together), but as previously noted do not show priming, as normals do, after a single prior presentation of an arbitrary, novel pairs of words. Such priming effects can be obtained after multiple presentations of the novel arbitrary word pair (Squire, 1992). Turning to animal studies, it is clear that some forms of classical or instrumental conditioning of responses to discrete salient cues are unaffected by hippocampal system damage (see O'Keefe & Nadel, 1978; Barnes, 1988; Rudy & Sutherland, 1994, for reviews). A fuller consideration of these forms of conditioning is presented in a later section.

(4) Lesions to the hippocampal system or bilateral electroconvulsive treatment (ECT) appear to give rise to a temporally graded retrograde amnesia for material acquired in the period of time preceding the lesion. Recent electrophysiological studies (Barnes, Jung, McNaughton, Korol, Andreasson, & Worley, 1994; Stewart & Reid, 1993) indicate that ECT has profound effects on hippocampal synapses. Although temporally-graded retrograde amnesia has been the subject of controversy (Warrington & Weiskrantz, 1978; Warrington & McCarthy, 1988), we believe the evidence is substantial enough to be taken seriously, and it plays a major role in the theory to be developed here. Early indications that retrograde amnesia may be temporally graded, at least in certain forms of amnesia, come from the observations of Ribot (1882) and from the early report of patient H.M. by Scoville and Milner (1957). More recent quantitative studies of a wide range of hippocampal amnesics suggests several conclusions (Squire, 1992):

- Hippocampal amnesics show a selective memory deficit for material acquired shortly before the date of their lesion. Memory for very remote material appears to be completely spared; in between there is an apparent gradient.

- The severity and temporal extent of the retrograde amnesia appears to vary with the extent of damage to the hippocampus and related structures.

[1] Jarrard (1993) treats the fact that Davidson, McKernan, and Jarrard (1993) find no effect of a lesion selective to the hippocampus *per se* in negative patterning as evidence against a role of the hippocampus in configural learning, but Rudy and Sutherland (1994) cite a total of 6 studies finding that selective hippocampal lesions lead to a deficit in negative patterning. Several of these studies use the ibotenate lesion of Jarrard (1989). Clearly the debate is not yet settled. Further discussion of the relationship between spatial and configural approaches may be found in the *General Discussion*.

- In some severe cases, the retrograde gradient can extend over periods of 15 years or more.

Results from animal studies are generally consistent with the human data, though in the case of the animal work the retrograde gradient appears to cover a much briefer span of time. Studies in rats (Winocur, 1990; Kim & Fanselow, 1992) have produced retrograde gradients covering a period of days or weeks. Primate experiments (Zola-Morgan & Squire, 1990) show a severe impairment relative to controls for memory acquired 2 or 4 weeks prior to surgery, but not for older memories.

A key observation is that there is a correspondence between the kinds of tasks that show retrograde amnesia and those that show anterograde amnesia. For example, Kim and Fanselow (1992) observed that the same rats who showed retrograde amnesia for the spatial context of a tone-shock association exhibited no retrograde amnesia for the simple tone-shock association itself.

A second crucial aspect of temporally graded retrograde amnesia is the fact that, after hippocampal lesions, performance on recent material can actually be worse than performance on somewhat older material. As Squire (1992) points out, this finding is crucial for the claim that some real consolidation takes place, since it rules out the alternative interpretation that memories are initially stored in two forms, whose effects are additive: a relatively transient, hippocampal-system dependent form, and a more persistent, hippocampal-system independent form. On this account, there is no alteration of the form of memory over time, there is merely decay. Nevertheless, because the decay of the hippocampal memory is more rapid, there would be a gradually diminishing difference between the two groups. Several animal studies now provide clear evidence against this simple dual-store interpretation (Zola-Morgan & Squire, 1990; Kim & Fanselow, 1992; Winocur, 1990); we show the data from all three in Figure 1. In all three studies, performance of lesioned animals at test is better when there is a longer delay between study and test, supporting a real change in the form or location of memory. Also shown are data from human ECT patients (Squire & Cohen, 1979), taken from a test called the *TV test* developed by Squire and Slater (1975). This test examined knowledge of single-season TV shows, for which memory depended primarily on exposure to the shows during the year they were aired. It is difficult to rule out the possibilities either that the depression in the years just prior to treatment affected initial storage. It also must be noted that the treatment may have affected more than just the hippocampal system. But no such difficulties apply to the findings from the animal studies, which are very clear in two of the cases: In both Winocur (1990) and Kim and Fanselow (1992), lesions occurring within 24 hours of the experience led to performance indistinguishable from chance, while lesions occurring at later points in time led to much better performance.

One Account of the Organization of Memory in the Brain

What follows is one account of the mechanisms of learning in the mammalian brain. The account is consistent with the data summarized above, and with several important anatomical and physiological findings that we will summarize below, and has many points in common with the accounts offered in several other synthetic treatments, beginning with Marr (1971). A comparison with these other treatments can be found in the *General Discussion*.

Our account begins with the assumption that the brain exploits complementary learning systems. One system relies on adaptation of synaptic connections among the neurons directly responsible for information processing and behavior. The other relies on adaptation of synaptic connections within a special memory system that includes the hippocampus and related structures.

The Neocortical Processing System

Adaptation of synaptic connections undoubtedly occurs in a wide range of neural processing systems in the brain, but for the cognitive forms of learning that are the principal focus of this paper, we will be concerned primarily with adaptive learning that is likely in most cases to occur in the neocortex. We suspect that the principles we propose for the neocortical system also apply to some other adaptive processing systems in the brain such as those that are involved in some forms of skill learning, including the basal ganglia and the cerebellum. We will comment in a later section on adaptive changes produced by animal conditioning paradigms in other systems such as the amygdala and various other sub-cortical brain structures.

We view the neocortex as a collection of partially overlapping processing systems, but for simplicity of reference we will speak of these systems collectively as a single system called “the neocortical processing system”. We include in this system those neocortical structures that we take to share the role of providing the neural substrate for higher-level control of behavior and cognitive processing, as well as other neocortical structures involved in sensory, perceptual, and output processes. Most but not all of the neocortex belongs to this system: the perirhinal and parahippocampal cortices are anatomically defined as neocortex, but they appear functionally to belong at least in part to the hippocampal memory system. It may be best to consider these as borderline areas in which the neocortical processing system and the hippocampal memory systems overlap. They certainly play a crucial role in mediating communication between the other parts of the hippocampal system and the neocortex.

We assume that performance of higher-level behavioral and cognitive tasks depends on the elicitation of patterns of activation over the populations of neurons in various regions of the

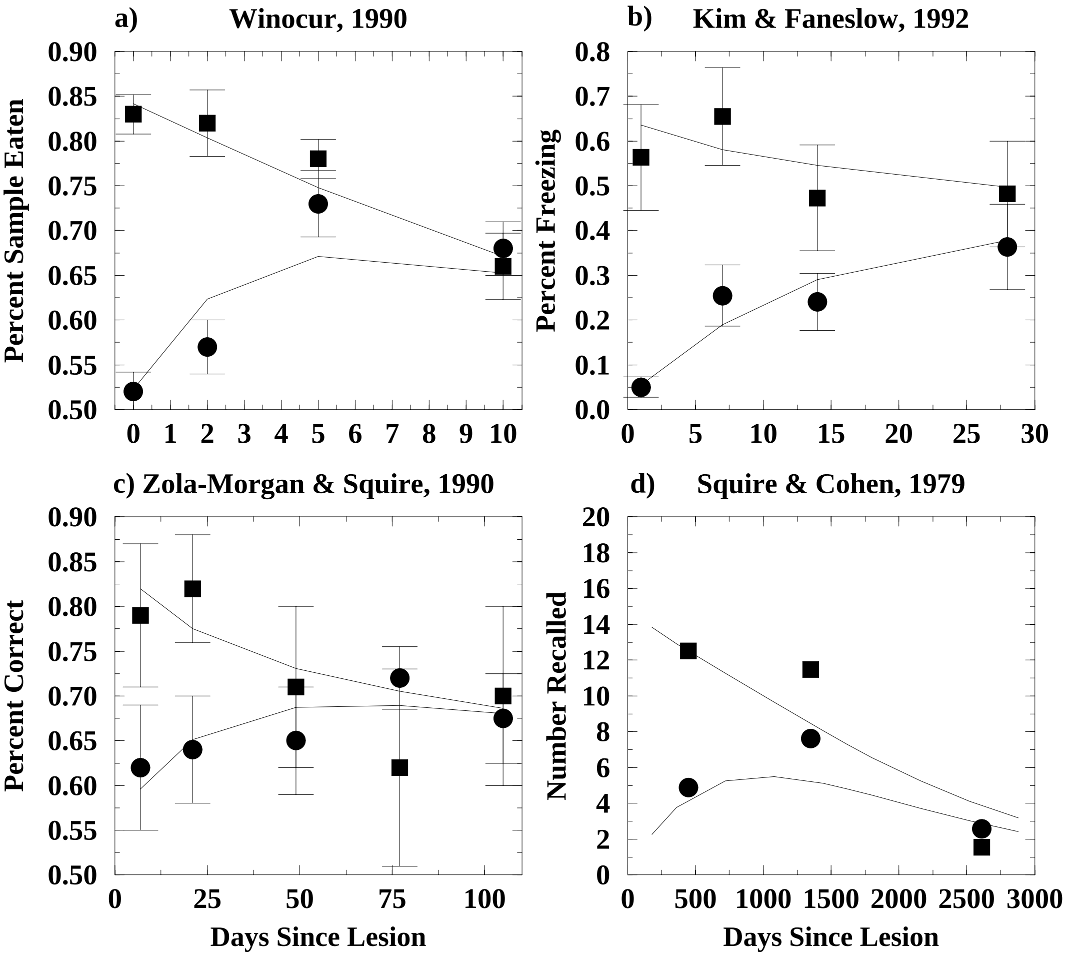

Figure 1: Panels (a) - (c) show behavioral responses of animals receiving extensive hippocampal system lesions (circles) or control lesions (squares) as a function of the numbers of days elapsing between exposure to the relevant experiences and the occurrence of the lesion. Bars surrounding each data point indicate the standard error. (a) Percent choice of a specific sample food (out of two alternatives) by rats exposed to a conspecific who had eaten the sample food. (b) Fear (freezing) behavior shown by rats when returned to an environment in which they had experienced paired presentations of tones with foot shock. (c) Choices of reinforced objects by monkeys exposed to 14 training trials with each of 20 object pairs. (d) Recall by depressed human subjects of details of television shows aired different numbers of years prior to the time of test, after electroconvulsive treatment (circles) or just prior to treatment (squares). Here we have translated years into days to allow comparison with the results from the animal studies. The curves shown on each panel are based on a simple model discussed in a later section and depicted in Figure 14, using the parameters shown in Table 1. *Note:* Data in (a) are from Figure 2 of “Anterograde and retrograde amnesia in rats with dorsal hippocampal or dorsomedial thalamic lesions”, by G. Winocur, 1990, *Behavioral Brain Research, 38*, p. 149. Copyright 1990 by Elsevier Science Publishers. Permission pending. Data in (b) are from Figure 2 of “Modality-specific retrograde amnesia of fear”, by J. J. Kim and M. S. Fanselow, 1992, *Science, 256*, p. 676. Copyright 1992 by the American Association for the Advancement of Science. Permission pending. Data in (c) are from Figure 2 of “The primate hippocampal formation: Evidence for a time-limited role in memory storage”, by S. Zola-Morgan and L. R. Squire, 1990, *Science, 250*, p. 289. Copyright 1990 by the American Association for the Advancement of Science. Permission pending. Data in (d) are from Figure 1 of “Memory and amnesia: Resistance to disruption develops for years after learning”, by L. R. Squire and N. Cohen, 1979, *Behavioral and Neural Biology, 25*, p. 118. Copyright 1979 by Academic Press, Inc. Permission pending.

neocortical system by other patterns of activation over the same or different regions. For example, in an acquired skill (such as reading), the pattern produced by an input (such as a printed word) elicits a corresponding pattern representing an output (such as the motor program for pronouncing the word). In a free association task, a pattern representing a stimulus word elicits another pattern representing the response word. In retrieving an arbitrary list-associate in a paired-associate learning task, the stimulus pattern must specify not only the stimulus word but also some information about the encoding context, but the principle remains the same: task performance occurs through the elicitation of one pattern of activation in response to another that serves as a cue. For this to work in tasks requiring the contextually-appropriate retrieval of patterns of activation representing specific propositions, events, etc., the system must be structured in such a way that any aspect of the content of the target pattern, as well as patterns representing material associated with the target pattern, can serve as retrieval cues.

Patterns are elicited by the propagation of activation via synaptic connections among the neurons involved. The knowledge that underlies the processing capabilities of the neocortex is stored in these connections. Thus the knowledge is assumed to be embedded in the the very neural circuits that carry out the tasks that use the information.

We assume that every occasion of information processing in the neocortical system gives rise to small, adaptive adjustments to the connections among the neurons involved. The adjustments are widely distributed across all of the relevant connections, but are very small in magnitude, and so have relatively subtle effects; they tend to facilitate a repetition of the same act of processing or an essentially similar one at a later time; and/or to facilitate reaching the same global state of activation (corresponding for example to an entire proposition or image) when given any fragment or associate of it as a cue.[2] We assume, however, that the changes that result from one or a few repetitions of an experience are not sufficient to support the reinstatement of a pattern representing a specific conjunction of arbitrary elements, such as the conjunction of an arbitrary pair of words in a paired-associate learning experiment, or the conjunction of elements that together compose a specific episode or event.

Over the course of many repetitions of the same or substantially similar acts of information processing, the changes to the synaptic connections among neurons in the neocortical system will accumulate. When the changes arise from the repetition of the same specific content, for example the association between a particular word and its meaning, the accumulation of such changes will provide the basis for correct performance in tasks that depend on the specific content in question. When they reflect different examples of some sort of structured relationship between inputs and outputs, such as, for example, the structured relation that holds between the spellings of words and their sounds, they will provide the basis of an acquired cognitive skill.

The Hippocampal Memory System

The representations of an experience in the neocortical system consist of widely distributed patterns of neural activity. As just noted, we assume that each experience gives rise to small adaptive changes, but that these will generally not be sufficient to allow rapid learning of arbitrary associative conjunctions that we assume provide the substrate for explicit recall of the contents of specific episodes and for other hippocampal-system dependent tasks. We assume that performance in such tasks depends initially on substantial changes to the strengths of connections among neurons in the hippocampal system. Information is carried between the hippocampal system and the neocortical system via bi-directional pathways that translate patterns of activity in the neocortical system into corresponding patterns in the hippocampal system and *vice versa*. We do not assume that the hippocampal system receives a direct copy of the pattern of activation distributed over the higher level regions of the neocortical system; instead the neocortical representation is thought to be re-represented in a compressed format over a much smaller number of neurons in the hippocampal system. McNaughton (1989) has referred to this compressed pattern as a "summary sketch" of the current neocortical representation. Such compression can often occur without loss of essential information if there is redundancy in the neocortical representations. The familiar data compression schemes that are used for computer files exploit such redundancy, and very high levels of compression may be possible if the patterns being compressed are highly constrained or redundant. Artificial neural networks structurally similar to those suggested by Figure 2 are quite commonly used to perform pattern compression and decompression (Ackley, Hinton, & Sejnowski, 1985; Cottrell, Munro, & Zipser, 1987). Compression is carried out in these models by the connections leading from the input to a much smaller representation layer, and decompression occurs via connections leading back from the representation layer to the input layer. Intermediate layers can be interposed on either the input or the output side to increase the sophistication of the compression and/or decompression processes. (For further discussion, see the section on *Binding* in the *General Discussion*.

Within the hippocampus itself, we assume that the event or experience is represented by a sparse pattern of activity, in which the individual neurons represent specific combinations or conjunctions of elements of the event that gave rise to the pattern of activation. We assume that once such a pattern of activity arises in the hippocampal memory system, it

[2] In general one expects adjustments of connection weights to produce a general facilitation of retrieval of the overall pattern through changes that occur among the neurons active in the retrieved pattern itself, as well as a more specific facilitation of retrieval from the same cue due to changes that occur between the neurons representing the retrieved pattern and those representing the cue.

may potentially become a stable memory. Plastic changes to the synapses on fibers coming into the hippocampus tend to increase the likelihood that a subsequent fragment of the pattern will elicit the entire pattern, and plastic changes to synaptic connections among the neurons active in the pattern tend to make this pattern an attractor—that is, a pattern toward which neighboring patterns or incomplete versions of the pattern will tend to converge. Several repetitions may be required for these changes to reach sufficient strength to subserve memory task performance. During recall, if a part of the pattern representing the episode arises again in the neocortical system, this will be translated into a part of the pattern corresponding to the previous event in the hippocampal memory system. If the input is sufficiently close to the stored pattern, and if the changes to the relevant synaptic efficacies were sufficiently large, this input would then lead the hippocampal memory system to tend to settle into the attractor, thereby filling in the missing aspects of the memory trace. The return pathways from the hippocampal system to the neocortex, together with pre-existing intracortical connections, then reverse the translation carried out by the forward connections, thereby completing the neocortical reinstatement of the event pattern and enabling appropriate overt responses. Reinstatement in such a system is assumed to be a matter of degree, varying with the adequacy of the probe, the amount of initial learning, subsequent interference and decay; and the sufficiency of a particular degree of pattern reinstatement for overt behavior will depend on the exact nature of the task and behavioral response required.

Reinstatement and Consolidation of Hippocampal Memories in the Neocortical System

As just described, reinstatement of patterns stored in the hippocampal memory system may occur in task-relevant situations, where the memory trace is needed for task performance. We assume that reinstatement also occurs in off-line situations, including active rehearsal, reminiscence, and other inactive states including sleep (Marr, 1971). In such cases, we assume that reinstatement in the hippocampal memory system gives rise, via the return connections, to reinstatement in the neocortical processing system. This would have two important consequences: First, reinstatement of the stored event in an appropriate context would allow the stored information to be used for controlling behavioral responses. Second, reinstatement would provide the opportunity for an incremental adjustment of neocortical connections, thereby allowing memories initially dependent on the hippocampal system gradually to become independent of it. To the extent that the hippocampal memory system participates in this reinstatement process, it can be viewed not just as a memory store but as the teacher of the neocortical processing system.

In our view, the same consolidation process applies to the development of a neocortical substrate for performance in semantic, encyclopedic, and episodic memory tasks. As we define these terms here, semantic memory tasks are simply those that require the use of information about categories and concepts, encyclopedic tasks require the use of specific factual information, and episodic memory tasks are those that require the use of information contained in a specific previous event or episode in which the subject was an observer or participant. In our view, there is no special distinction between such tasks. Performance in all three depends initially on plastic changes within the hippocampal system, but the knowledge underlying all three can eventually be stored in the neocortical system via the gradual accumulation of small changes. Let us consider the specific episode or event in which one first encounters some particular fact: For example, that Neil Armstrong uttered the words “That’s one small step for [a] man …” when he first set foot on the moon. Such factual information would be encountered first in a particular context, in this case, perhaps, in the context of watching Armstrong live on TV as he set foot on the moon, during a family reunion celebrating grandfather’s 70th birthday. If the event of Armstrong’s landing is reinstated repeatedly, the accumulated changes to neocortical connections could eventually come to preserve the common aspects of the reinstated event. The result would allow the individual to perform correctly in the encyclopedic memory task of recalling what Armstrong said. If the previous reinstatements had specifically included information about the time and place of initial learning, then this information, too, would gradually become incorporated into the connection weights in the neocortical system, and would sustain performance in an episodic memory task. If, however, the reinstatements occur in many different contexts, and if these reinstatements do not include other aspects of the original context of encoding, no reliable memory for any particular context would remain. Much the same process would also apply to the learning of semantic information, such as the fact that giraffes have long necks, or the fact that a particular category label is the correct name to apply to a set of items derived from the same prototype. Knapp and Anderson (1984) and McClelland and Rumelhart (1985) both present connectionist models in which semantic memory and category learning can arise from the gradual accumulation of small changes resulting from individual events and experiences.

Evidence and Comment

Our accounts of neocortical processing and learning, of hippocampal involvement in some forms of memory, and of reinstatement of hippocampal memories during off-line periods are all grounded in evidence from neuroanatomical and neurophysiological investigations. We discuss the evidence for each of these aspects of our account in turn. There are certainly gaps but we do not dwell on these—we simply describe the evidence that is available.

Neocortical processing and learning. The basic notion that information processing takes place through the propagation of activation among neurons via synaptic connections does not appear to be in dispute, and is based on over 100 years of neurophysiological investigation. The evidence also strong that the neocortical processing system consists of a large number of interconnected brain areas. A compelling example is the parallel and hierarchical organization of the visual system (Felleman & Van Essen, 1991). Neurons in each area project to other neurons, both within the same area and in several other areas. Activation is propagated into and out of this system of brain regions via a number of different pathways specific to particular sensory modalities and/or effector systems. Generally, the projections between brain areas are bi-directional, so that activity in one part of the system can potentially influence activity in many other parts, in both feed-forward and feed-back directions.

The idea that processing depends on the pattern of synaptic connections among neurons and that adaptive changes in processing occur through the modification of such connections is axiomatic in neuroscience. The notion that higher-level, cognitive forms of learning are mediated by plastic changes in these connections goes back at least to James (1890) and Hebb (1949). At a physiological level, these changes probably occur through strengthening and weakening, as well as creation and pruning, of synaptic contacts between neurons. There is strong evidence of changes in functional connectivity in neocortex as a consequence of experience (Greenough, Armstrong, Cummery, Hawry, Humphreys, Kleim, Swain, & Wang, 1994; Gilbert, 1994; Merzenich, Recanzone, Jenkins, & Grajski, 1990; Singer & Artola, 1994; Kaas, 1994). Much of this work, is, however, restricted to primary sensory systems, and it remains to be determined exactly what the signals are that govern the plastic changes. Although neocortical synapses exhibit experience-dependent plasticity (Lee, 1983), the bulk of our understanding of synaptic plasticity comes from physiological studies in the hippocampal system.

Hippocampal involvement in some forms of memory. Our account requires that patterns of activity are propagated into and out of the hippocampal system during information processing. Neuroanatomically, it is clear that the necessary reciprocal pathways exist to perform these proposed functions (see Figure 2 from Squire, Shimamura, & Amaral, 1989b). The entorhinal cortex is the final convergence zone for neocortical inputs to the hippocampus, and is the main structure mediating return projections from the hippocampus to the neocortex. The entorhinal cortex is quite small relative to the combined size of all the neocortical regions that project to it, suggesting the need for a high degree of data compression as previously discussed. As the figure indicates, some of the the neocortical inputs to the entorhinal cortex and return pathways from the entorhinal cortex are mediated by the parahippocampal and perirhinal cortices, which may serve as the intermediate layers in a sophisticated compression/decompression operation.

Our account also asserts that the hippocampal system makes use of representations that are highly specific to the particular conjunctions or combinations of inputs that are active in particular experiences. A growing body of data from single unit recording in rats is consistent with the idea that these representations arise in the hippocampus itself. Part of this evidence comes from recordings of neurons in the CA3 and CA1 regions of the hippocampus in spatial tasks, for example a task in which the animal explores an eight-arm, radial maze to find food rewards at the ends of the arms. Neurons in these regions fire in a highly selective manner exhibiting ‘place fields’ (O’Keefe & Dostrovsky, 1971; O’Keefe & Conway, 1978), in contrast to neurons in the entorhinal cortex. Though spatially tuned to some extent, entorhinal neurons tend to fire much less selectively. This is illustrated in Figure 3 from McNaughton and Barnes (1990), which contrasts the spatial firing pattern of a typical neuron in CA3 with that of a typical entorhinal neuron, based on data from Barnes, McNaughton, Mizumori, Leonard, and Lin (1990). Place fields of hippocampal neurons can be seen as representing conjunctions of cues (including cues arising from the animal’s inertial sense of direction and location, Knierim, Kudrimoti, & McNaughton, in press) that go together to define a place in the environment. In fact, it may be better to think of these neurons as coding for conjunctions that define *situations* rather than just places, since the firing of a hippocampal neuron in a particular location in space is conditional on the task as well as the animal’s location in the environment (Qin, Markus, McNaughton, & Barnes, 1994; Gothard, Skaggs, Moore, & McNaughton, 1994). Also, note that the hippocampal representation is very sparse compared to the entorhinal input. In a given situation, a far smaller percentage of neurons in the hippocampus itself are firing than in the entorhinal cortex (Barnes et al., 1990; Quirk, Muller, & Kubie, 1990). The use of sparse, conjunctive coding in the hippocampus means that its representations of situations that differ only slightly may have relatively little overlap (Marr, 1969; McNaughton & Morris, 1987; O’Reilly & McClelland, 1994).

Our account requires the availability of a mechanism for synaptic plasticity in the hippocampus, and specifically assumes that the synaptic changes provided by these changes serve as the substrate of initial learning in hippocampal-system dependent memory tasks. There is now considerable evidence that such a mechanism exists in the hippocampus (see McNaughton & Morris, 1987; McNaughton & Nadel, 1990, for reviews). The mechanism is a form of plasticity known as associative long-term potentiation (LTP). LTP has been studied extensively in the rodent hippocampus since the studies of Bliss and Gardner-Medwin (1973) and Bliss and Lømo (1973). Associative LTP is found particularly in the synapses from the axons of the principal neurons of the entorhinal cortex onto the dendrites of the principal neurons in the dentate and CA3 regions of the hippocampus, as well as the synapses from the axons of the principal neurons in CA3 onto the dendrites of other neurons in CA3 and neurons in CA1. LTP may be the

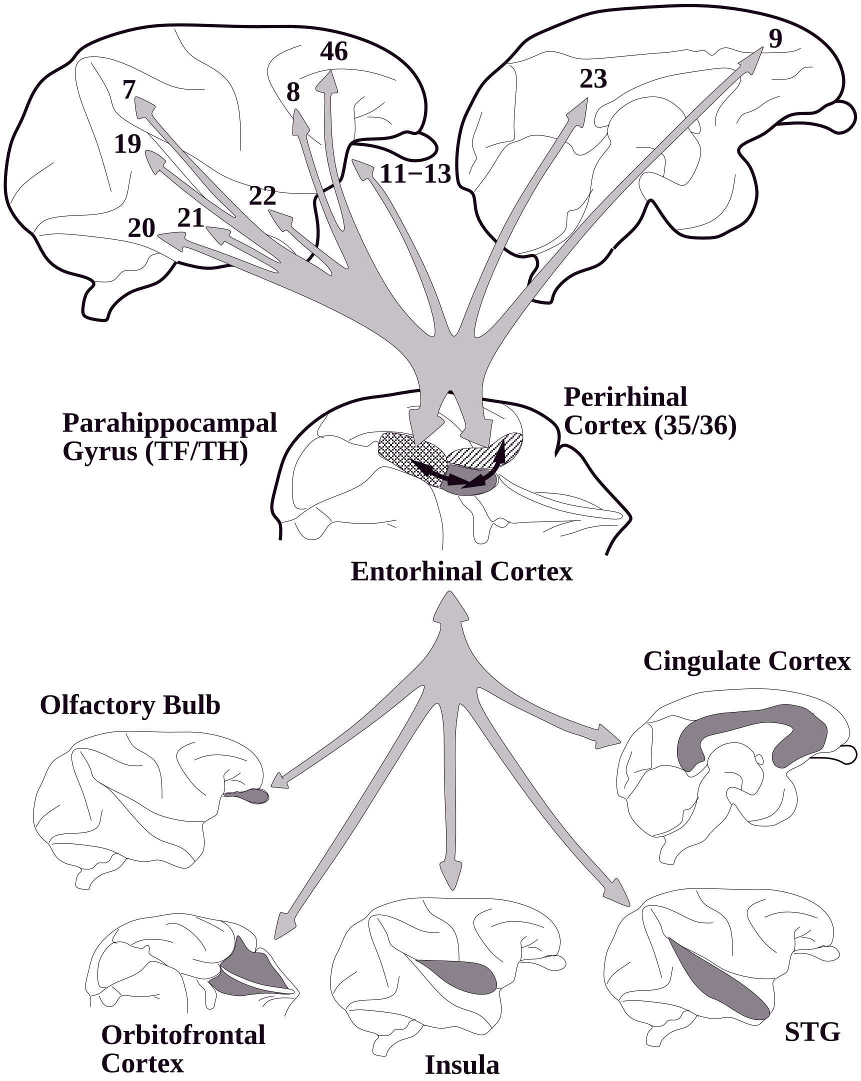

Figure 2: Schematic representation of the inputs and outputs of the hippocampal system in the primate. The upper panel shows the frontal, temporal, and parietal areas reciprocally connected with the parahippocampal gyrus and the perirhinal cortex, which in turn connect reciprocally with the entorhinal cortex. The lower panel shows the areas that have direct reciprocal connections with the entorhinal cortex. *Note*: From Figure 10 of "Memory and the Hippocampus" (p. 227), by L. R. Squire, A. P. Shimamura, and D. G. Amaral, in *Neural Models of Plasticity: Experimental and Theoretical Approaches*, edited by J. H. Byrne and W. O. Berry, 1989, New York: Academic Press. Copyright 1989 by Academic Press, Inc. Reprinted with permission.

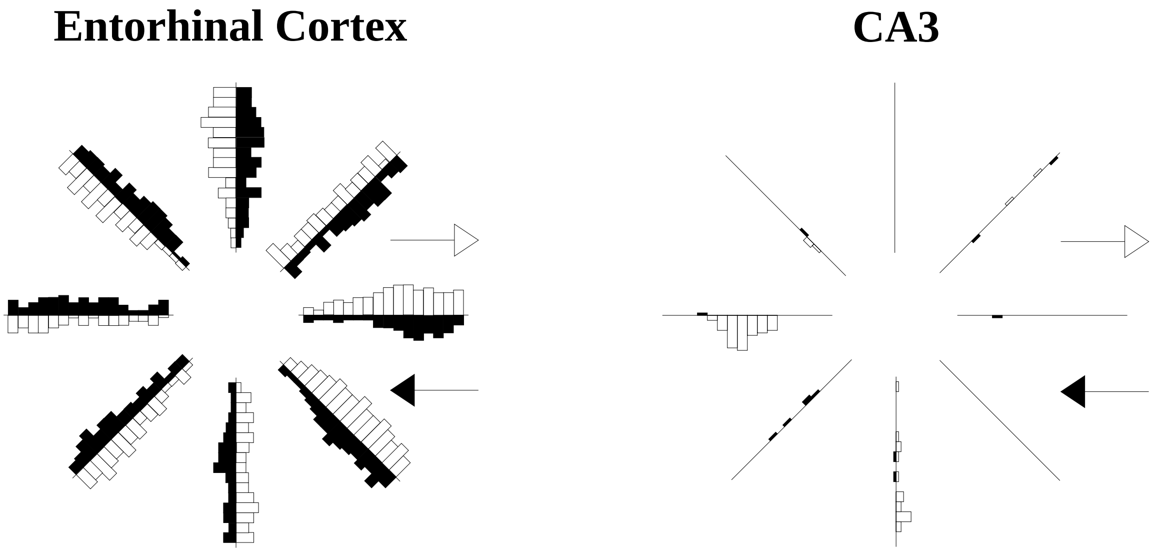

Figure 3: Response profiles of a representative neuron in area CA3 of the hippocampus and of a representative neuron in entorhinal cortex during performance in a spatial working memory task in the 8-arm radial maze Response profiles are shown in the form of bars perpendicular to the arms of the runway, indicating the number of times the particular neuron fired as the animal repeatedly traversed the maze. Black bars indicate firing as the animal progressed inward on the indicated arm, and white bars indicate firing as the animal progressed outward. *Note:* From Figure 4 of "From cooperative synaptic enhancement to associative memory: Bridging the abyss" by B. L. McNaughton and C. A. Barnes, 1990, *Seminars in the Neurosciences, 2*, p. 412. Copyright 1990 by W. B. Saunders Company. Reprinted with permission.

experimental manifestation of the synaptic modifications that underlie the contribution of the hippocampus to some forms of learning and memory (Marr, 1971; McNaughton & Morris, 1987). LTP in these synapses can last for days to weeks depending on the intensity and duration of the inducing stimulation (Barnes, 1979; Abraham & Otani, 1991), and is associative in that it normally depends on near-synchronous input to the receiving neuron by a large number of convergent fibers (McNaughton, Douglas, & Goddard, 1978; Levy & Steward, 1979; Barrionuevo & Brown, 1983). This form of LTP depends critically on the activation of a special type of post-synaptic receptor known as the NMDA receptor, which is activated by the conjunction of transmitter release from the pre-synaptic terminal and post-synaptic depolarization (Wigström, Gustaffson, & Huang, 1986). Chemical agents that block the NMDA receptor prevent long-term potentiation with little effect on the transmission process *per se* (Collingridge, Kehl, & McLennan, 1983; Harris, Ganong, & Cotman, 1984). Crucially, these agents also block learning in spatial memory tasks without interfering with expression of learning that occurred at a previous time or producing retrograde amnesia (Morris, Anderson, Lynch, & Baudry, 1986). In addition, electrical stimulation that sufficiently saturates LTP in the hippocampus also produces profound deficits in spatial learning (Barnes et al., 1994) and a temporally limited retrograde amnesia (McNaughton, Barnes, Rao, Baldwin, & Rasmussen, 1986).

As previously noted, there is evidence from lesion studies of involvement of parts of the hippocampal system other than the hippocampus itself in several forms of learning. Lesions including these structures can produce deficits that are more severe than lesions restricted to the hippocampus, suggesting that some of the plastic changes underlying performance in some hippocampal-system dependent tasks may lie outside the hippocampus itself. Exactly how plasticity in other zones contributes to memory performance is a matter of considerable ongoing discussion (see Eichenbaum et al., 1994, and the accompanying commentary). There are several possibilities that are consistent with our overall account. First, particular subareas of the parahippocampal region may be involved in bidirectional communication of specific types of information between the hippocampal system and the neocortex; if so we would expect lesions to these sub-areas to have specific effects on memories that involve the relevant types of information, as Suzuki (1994) suggests. Second, as a borderline area between hippocampus and neocortex, the parahippocampal region may participate is some forms of information processing, including, for example, retention of information about the recent occurrence of novel stimuli for short periods of time (Gaffan & Murray, 1992). Such functions may co-exist with these areas' involvement in bi-directional communication between the hippocampus itself and the rest of the neocortical system. A third possibility is that there is a hierarchy of plasticity, such that learning in the hippocampus itself is very rapid, learning in the neocortical system is very slow, and learning within the parahippocampal region occurs at an intermediate rate. This could explain why damage restricted to the hippocampus itself may produce a milder deficit in new learning and a milder retrograde amnesia than more extensive hippocampal system lesions (Zola-Morgan, Squire, & Amaral, 1986; Squire, Zola-Morgan, & Alvarez, 1994).

Reinstatement of hippocampal memories. There is little direct evidence of hippocampal involvement in the reinstatement of patterns of activity in the neocortex, but there is evidence of reinstatement in the hippocampus itself of patterns derived from recent experiences. The evidence is based on activity recorded in rats during periods of slow wave sleep. In both primates and rodents, hippocampal electrical activity during slow wave sleep (as well as during quiet wakefulness) is characterized by a unique pattern called sharp waves (O'Keefe & Nadel, 1978; Buzsaki, 1989). Hippocampal sharp waves are brief periods of quasi-synchronous, high-frequency burst discharge of hippocampal neurons, lasting about 100 msec. In theory, such activity provides the optimal conditions for synaptic plasticity in downstream neurons (Douglas, 1977; McNaughton, 1983; Buzsaki, 1989). Buzsaki (1989) and his colleagues (Chrobak & Buzsaki, 1994) have provided a strong case that sharp waves arise in hippocampal area CA3 and are propagated both to area CA1 and to the output layers of the entorhinal cortex, from which they could be propagated widely to the neocortical system. Thus, patterns stored in the hippocampus might complete themselves during hippocampal sharp waves, thereby providing an opportunity for reinstatement in the neocortex. Simulation studies (Shen & McNaughton, 1994) demonstrate that random activity can sometimes lead to reinstatement of attractors previously stored in a hippocampus-like associative network. In support of these proposals, Pavlides and Winson (1989) have shown that hippocampal neurons which have been selectively activated during a prior episode of waking behavior are selectively more active during subsequent slow wave and paradoxical sleep. More recently, Wilson and McNaughton (1994b, 1994a) have found that the cross-correlation structure that arises in a large population (50-100) of simultaneously recorded CA1 neurons during exploration of an environment is preserved in subsequent sharp-wave activity while the animal is resting or sleeping in an entirely different apparatus. This correlational structure is absent during sleep periods before exploration. Thus, there is now strong empirical support for the idea that memory traces—or at least, correlations of activity associated with such traces—are indeed reactivated in the rat hippocampus during "off-line" periods.

Summary

We can now summarize our account of the organization of the memory system by noting how it accounts for the main features of the pattern of deficits and spared performance found following a hippocampal system lesion:

The deficit in the ability to learn new arbitrary associations involving conjunctions of cues from a few exposures would arise from the fact that these would have been stored in the (now destroyed) hippocampal system; the small changes that would occur in the neocortical system could contribute to repetition priming effects but would be insufficient to support normal rates of acquisition in semantic and episodic memory tasks and other tasks that depend on the acquisition of novel conjunctions of arbitrary material.

The spared acquisition of skills would arise from the gradual accumulation of small changes in the connections among the relevant neural populations in the neocortical system as well as other relevant brain systems. The temporally extended and graded nature of retrograde amnesia would reflect the fact that information initially stored in the hippocampal memory system can become incorporated into the neocortical system only very gradually, due to the small size of the changes made on each reinstatement.

The ability of even very profound amnesics to acquire often-repeated material gradually (Milner et al., 1968; Glisky, Schacter, & Tulving, 1986a, 1986b) would likewise reflect this slow accumulation of changes in the neocortical system after the onset of amnesia. The fact that such learning is often restricted to the specific task contexts in which it was acquired follows from the assumption that the learning actually takes place directly within the connections among the neural populations that were activated during the acquisition process.

Our account of the organization of learning in the brain is intended as a provisional factual characterization. It embodies some unproven assumptions, and so it might be viewed as a theory of memory in some sense. We offer it, however, not as a theory in itself, but as a starting place for theoretical discussion. Though it is neither fully explicit nor complete (some gaps, such as a consideration of spared conditioning of responses to individual salient cues, will be discussed in later sections), the account appears to be broadly compatible with a large body of data, and it is consistent enough with many of the other accounts considered in the general discussion that we suggest it is useful to treat it as provisionally correct, at least in its essentials.

Key Questions about the Organization of Memory in the Brain

Supposing provisionally that our account is basically correct, we can now ask, why is it that the system is organized in this particular way?

Two key functional questions arise:

- Why do we need a hippocampal system, if ultimately performance in all sorts of memory tasks depends on changes in connections within the neocortical system? Why are the changes not made directly in the neocortical system in the first place?

- Why does incorporation of new material into the neocortical system take such a long time? Why are the changes to neocortical connections not made more rapidly, shortly after initial storage in the hippocampal system?

Successes and Failures of Connectionist Models of Learning and Memory

The answers we will suggest to these questions arise from the study of learning in artificial neural network or connectionist models that adhere to many aspects of the account of the mammalian memory system given above, but which do not incorporate a special system for rapid acquisition of the contents of specific episodes and events. Such networks are similar to the neocortical processing system, in that they may consist of several modules and pathways interconnecting the modules, but they are monolithic in the sense that knowledge is stored directly in the connections among the units of the system that carries out information processing, and there is no separate system for rapid learning of the contents of particular inputs.

Discovery of Shared Structure through Interleaved Learning

The first and perhaps most crucial point is that in such monolithic connectionist systems there are tremendous ultimate benefits of what we will call *interleaved learning*. By interleaved learning we mean learning in which a particular item is not learned all at once, but instead is acquired very gradually, through a series of presentations interleaved with exposure to other examples from the domain. The adjustments made to connection weights on each exposure to an example are small so that the overall direction of connection adjustment is governed, not by the particular characteristics of individual associations, but by the shared structure common to the environment from which these individual associations are sampled.

Consider, in this context, some of the facts we know about robins. We know that a robin is a bird, it has wings, it has feathers, it can fly, it breathes, it must eat to stay alive, and so on. This knowledge is not totally arbitrary knowledge about robins but is in fact part of a system of knowledge about robins, herons, eagles, sparrows and many other things. Indeed much of the information we may have about robins probably does not come from specific experience with robins but from other, related things. Some such knowledge comes from very closely related things of which we may have knowledge, such as other birds; while other knowledge may come from other things less closely related but still related enough in some particular ways to support some knowledge sharing, such as other animals, or even other living things. A key issue for our use of concepts is the fact that what counts as related is by no means obvious, and is not in general predictable from surface properties. Birds are more related to, for example, reptiles and fish than they are to insects.

Connectionist models that employ interleaved learning suggest how knowledge of relations among concepts may develop. Both Hinton (1989) and Rumelhart (1990; Rumelhart & Todd, 1993) developed simulations to illustrate how connectionist networks can learn representations appropriate for organized bodies of conceptual knowledge. We use the Rumelhart example here, because it relates to the domain of knowledge about living things that we have already begun to consider as an example, and because, as we shall see, there is some empirical data about the development of children's knowledge that this model can help us understand. The specific example is highly simplified and abstract. It captures approximately the constraints that may be operative in the discovery of conceptual structure from an ensemble of sentences that convey simple propositional statements about living things, in that concepts are represented by arbitrary tokens (akin to words) rather than by percepts that directly provide some information about the concepts under consideration. The conceptual structure resides not in the direct appearance of the words that convey the concepts but in the relations that the concepts referred to by the words enter into with other concepts.

Human knowledge of the domain of living things appears to be organized hierarchically, with a principal grouping into plants and animals, and then other, finer, subgroupings within each of these broad classes (we refer not to objective biological information *per se* but to the cognitive representations that people have of this information). Previous, symbolic approaches to knowledge representation directly imported the hierarchical organization of knowledge into their structure, representing knowledge about concepts in a data structure known as a *semantic network* (Quillian, 1968; see Figure 4). Such networks are not to be confused with connectionist networks, since they represent and process information in fundamentally different ways. In the semantic network, concepts are organized hierarchically, using links called *isa* links, as a short form of the statement *An X is a Y*. Given this organization, semantic networks could store knowledge of concepts in a succinct form, with information that is true of all of the concepts in an entire branch of the tree at the top of the branch. For example, the predicate *has feathers* can be stored at the *bird* node, since it is true of all birds. This allows generalization to new instances. When a new type of thing is encountered, for example an egret, we need only to be told that it is a bird, and to link the an new node for *egret* to the node for *bird* by an *isa* link. Then our knowledge of egrets can inherit all that is known about birds.

Semantic networks of this type were very popular vehicles for representation for a period of time in the 1970's, but apparent experimental support (Collins & Quillian, 1969) for the hypothesis that people's knowledge of concepts is organized this way was illusory (Rips, Shoben, & Smith, 1973). Com-

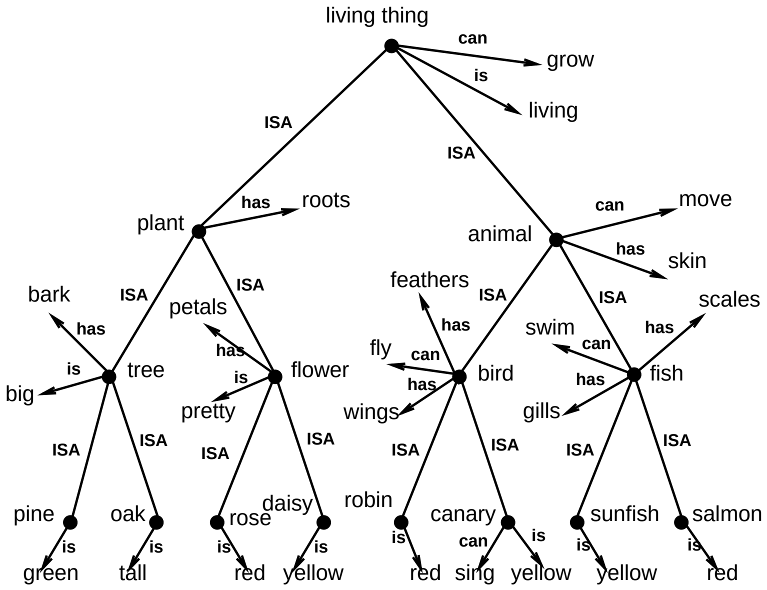

Figure 4: A semantic network of the type formerly used in models of the organization of knowledge in memory. All of the propositions used in training the network are based on the information actually encoded in this figure. For example the network indicates that living things can grow; that a tree is a plant; and that a plant is a living thing. Therefore it follows that a tree can grow. All of these propositions are contained in the training set. *Note:* Redrawn with alterations from Figure 1.8 of “Learning and connectionist representations”, (p. 14), by D. E. Rumelhart and P. M. Todd, in *Attention and Performance XIV: Synergies in Experimental Psychology, Artificial Intelligence, and Cognitive Neuroscience,* edited by D. E. Meyer and S. Kornblum, (1993), Cambridge, MA: MIT Press. Copyright 1993 by MIT Press. Permission pending.

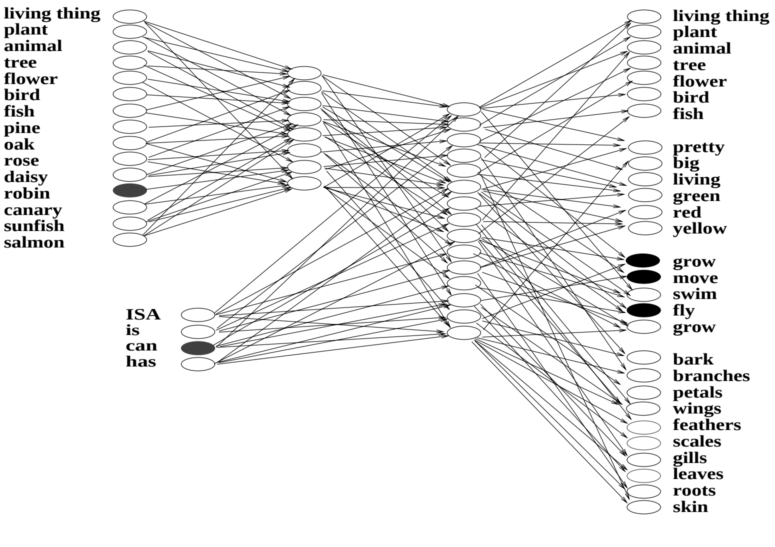

Figure 5: Our depiction of the connectionist network used by Rumelhart to learn propositions about the concepts shown in Figure 4. The entire set of units used in the actual network is shown. Inputs are presented on the left, and activation propagates from left to right. Where connections are indicated, every unit in the pool on the left (sending) side projects to every unit in the right (receiving) side. An input consists of a concept-relation pair; the input *robin can* is illustrated here by darkening the active input units. The network is trained to turn on all those output units that represent correct completions of the input pattern. In this case, the correct units to activate are *grow*, *move* and *fly*; the units for these outputs are darkened as well. Subsequent analysis focuses on the concept representation units, the group of eight units to the right of the concept input units. Based on the network depicted in Rumelhart and Todd (1993), Figure 1.9, page 15.

putationally, semantic networks of this type become cumbersome to use when they contain a large amount of information (Fahlman, 1981). It becomes very difficult to determine when it is appropriate to consider a property to be essentially common to a category even though there are exceptions, and when it is appropriate to consider a property sufficiently variable that it must be enumerated separately on the instances. The problem is compounded by the fact that most concepts are constituents of multiple intersecting hierarchies, in which case intractable inheritance conflicts can arise.

Connectionist models offer a very different way of accounting for the ability to generalize knowledge from one concept to another. According to this approach (Hinton, 1981; Touretzky & Geva, 1987), generalization depends on a process that assigns each concept an internal representation that captures its conceptual similarity to other concepts. This alternative approach appears to be more consistent with the psychological evidence (Rips et al., 1973), since the evidence favors the view that conceptual similarity judgments are made by comparing representations of concepts directly, rather than searching for common parents in a hierarchically structured tree. This alternative also overcomes the vexing questions about how to handle partially regular traits and exceptions, since idiosyncratic as well as common properties can be captured in these representations.

The approach depends on exploiting the ability of a network to discover the relations among concepts through interleaved learning. The network is trained on a set of specific propositions about various concepts, and in the course of training, it learns similar representations for similar concepts. By similar concepts, we mean concepts that enter into overlapping sets of propositions.

Rumelhart trained a network on propositions about a number of concepts: living things, plants, animals, trees, oaks, pines, flowers, roses, daisies, animals, birds, canaries, robins, fish, salmon, and sunfish. The training data were the set of true propositions either explicitly represented in or derivable from the semantic network shown in Figure 4. The connectionist network used to learn these propositions is shown in Figure 5. It consists of a number of nonlinear connectionist processing units organized into several modules, connected as illustrated in the figure. Where arrows are shown they signify complete connectivity from all the units in the module at the sending end of the arrows to all of the units at the receiving end.

Input to the network is presented by activating the unit for a concept name in the concept input module on the upper left, and the unit for a relation term in the relation input module on the lower left. The relations *isa*, *has*, *can* and *is* are represented. The task of the network is to respond to each input by activating units in the appropriate module on the right corresponding to the correct completion or completions of the input. For example in the case of the input *robin isa* the network is trained to activate the output units for *living thing*, *animal*, *bird*, and *robin*. In the case of *robin can* the network is trained to activate the output units for *grow*, *move*, and *fly*. The inputs and desired outputs for this latter case are indicated in the figure.

Before learning begins, the network is initialized with random weights. At first when an input is presented, the output is random and bears no relation to the desired output. The goal is to adjust these connection weights, through exposure to propositions from the environment, so as to minimize the discrepancy between desired and obtained output over the entire ensemble of training patterns. This goal can be achieved by interleaved learning using a gradient descent learning procedure: During training, each pattern is presented many times, interleaved with presentations of the other patterns. After each pattern presentation, the error—i.e., the discrepancy between desired and obtained output—is calculated. Each connection weight is then adjusted either up or down by an amount proportional to the extent that its adjustment will reduce the discrepancy between the correct response and the response actually produced by the network. The changes to the connection weights are scaled by a learning rate constant ϵ that is set to a small value, so that only small changes are made on any given training trial. Thus, responses are learned slowly. Over time, some of the changes made to the connections are mutually cooperative and some of the changes cancel each other out. The cooperative changes build up over time, with the end result that the set of connections evolves in a direction that reflects the aggregate influence of the entire ensemble of patterns.

To understand the results of the cooperative learning, we will consider patterns of activation the network comes to produce on the eight units in the module to the right of the concept units in the figure. These units are called the *concept representation units*. The patterns of activation in this module can be considered to be the learned internal representations of each concept; the connections from the concept input units to the representation units can be viewed as capturing the mapping between input patterns and internal representations. The rest of the connections in the network can be seen as capturing the mapping from these internal representations, together with patterns on the relation units, to appropriate response patterns at the output layer.

In the course of learning, the network learns both how to assign useful representations, and how to use these to generate appropriate responses. That is, it learns a set of input-to-representation weights that allow each concept to activate a useful internal representation, and it learns a set of weights in the rest of the network that allows these representations to produce the correct output, conditional on this representation and the relation input. Note that there is no direct specification of the representations that the network should assign; the representations—and the connection weights that produce them—arise as a result of the action of the learning procedure.

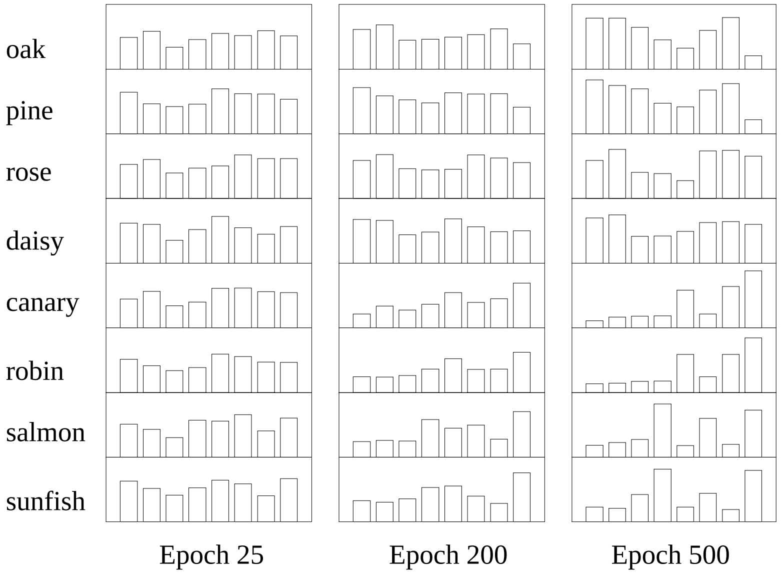

Figure 6: Representations discovered in our replication of Rumelhart's learning experiment, using the network shown in Figure 5. The figure presents a vertical bar indicating the activation of each of the eight concept representation units produced by activating the input unit for each of the eight specific concepts. The height of each vertical bar indicates the activation of the corresponding unit on a scale from 0 to 1. One can see that initially all the concepts have fairly similar representations. After 200 epochs, there is a clear differentiation of the representations of the plants and animals, but the trees and flowers are still quite similar as are the birds and the fish. After 500 epochs, the further differentiation of the plants into trees and flowers and of the animals into fish and birds is apparent.

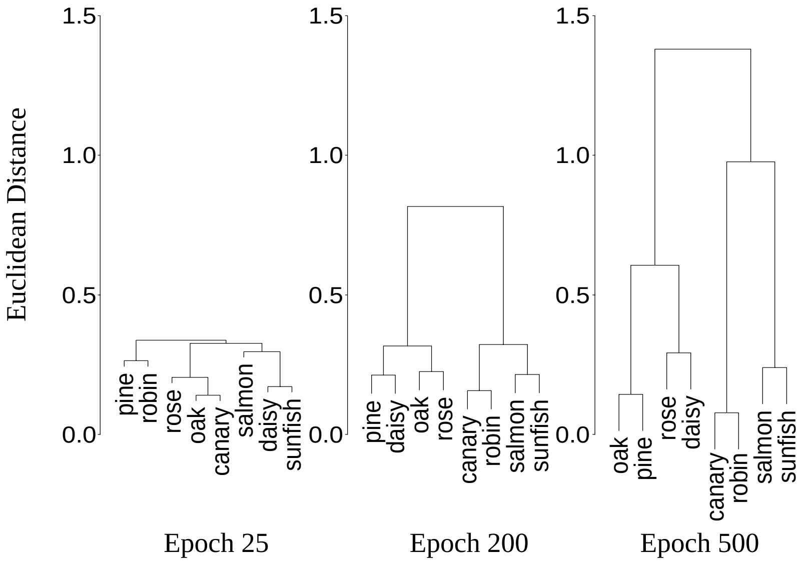

Figure 7: Similarity structure discovered in our replication of Rumelhart's learning experiment, using the representations shown in Figure 6. These analyses make the similarity relationships among the patterns shown in the preceding Figure explicit. The clustering algorithm recursively links a a pattern or an previously-linked group of patterns to another pattern or previously formed group. The process begins the the pair that is most similar, the elements combined are then replaced by the resulting group, and the process continues until everything have been joined into a single superordinate group. Similarity is measured by the Euclidian distance metric (Sum of the squared differences between the activations of the corresponding elements in the two patterns). The height of the point where a subtree branches indicates the Euclidean distance of the elements joined at that branch point.

We repeated Rumelhart's simulations, training the network for a total of 500 epochs (sweeps through the training set) using the gradient descent learning procedure.[3] The representations at different points in training are shown in Figure 6. These are simply the patterns of activation over the representation units that arise when the input unit corresponding to each of the eight specific concepts is activated. The arrangement and grouping of the representations, shown in Figure 7, reflects the similarity structure among these patterns, as determined by a hierarchical clustering analysis using Euclidian distance as the measure of similarity of two patterns. At an early point in learning (Epoch 25), the analysis reveals an essentially random similarity structure, illustrating that at first the representations do not reflect the structure of the domain: For example *oak* is grouped with *canary* indicating that the representation of *oak* is more similar at this point to *canary* than it is to *pine*. At later points in training, however, the similarity structure begins to emerge. At Epoch 500, we see that the complete hierarchical structure is apparent: The two trees (*oak* and *pine*) are more similar to each other than either is to any other concept, and the representations of the two flowers, the two birds, and the two fish are more similar to each other than either member of any of these pairs is to the representation of any other concept. Furthermore, the representations of the trees are more similar to the representations of the flowers than they are to the representations of any of the animals, and the representations of the birds are more similar to the representations of the fish than they are to the representations of any of the plants. Examination of the clustering of the representations at Epoch 200 shows that the network has by this point only learned the coarser distinction between plants and animals, since at this point the plants and animals are well differentiated but within the plants and animals the differences are very small and not yet completely systematic with respect to subtype. For example, *pine* is grouped with *daisy* rather than *oak*. Thus we see that the network exhibits a progressive differentiation of concepts, progressing from coarser to finer conceptual distinctions through the course of learning.

The similarity structure shown in Figure 7—for example, the fact that *oak* and *pine* are similar to each other but quite different from *canary* and *robin*—arises not because of intrinsic similarity among the inputs, but because of similarity among the responses the network must learn to make when the various concepts are presented with the same relation term. The connections in the rest of the network exploit these similarities, so that what the network has learned about one concept tends to transfer to other concepts that use similar representations. We can illustrate this by examining what happens if, after training on the material already described, a new concept is added such as *sparrow*, and the network is taught only the correct response to the *sparrow isa* input, interleaving this example with the rest of the training corpus (Rumelhart, 1990 performed a very similar experiment). Through this further training, the network assigns a representation to *sparrow* that is similar to the representation for *robin* and *canary*. This allows correct performance, since such a representation is already associated with the correct output for the *isa* relation term. This representation is also already associated with the correct responses to be made when it is active in conjunction with the other relation terms. Therefore the network will respond appropriately when the other relation terms are paired with *sparrow*, even though it has never been trained on these cases. In fact the network correctly sets the activity of all those outputs on which *canary* and *sparrow* agree; where they disagree it produces compromise activations reflecting the conflicting votes of the two known bird concepts.

The ability to learn to represent concepts so that knowledge acquired about one can be automatically shared with other related concepts is, we believe, a crucial cognitive capacity that plays a central role in the very gradual process of cognitive development. The order of acquisition of conceptual distinctions in such systems, beginning with coarser distinctions and proceeding to finer distinctions between subtypes, mirrors the developmental progression from coarser to finer distinctions studied by Keil (1979). Keil was interested in the conceptual differentiation of children's knowledge of different kinds of things, not so much in terms of the specific facts they knew about them, but in terms of the range of things that they believed could plausibly be said about them, or in Keil's terms *predicated* of them. As adults, we know, for example, that it is appropriate to attribute a duration to an event (such as a lecture or movie), but not to an animate being or physical object (such as a person or a chair). Feelings, on the other hand, can be attributed to humans, but not to plants or inanimate objects. Thus we can predicate a duration to an event and a feeling to a person, but we cannot predicate a duration to a person or a feeling to an event. To assess children's knowledge of these matters, Keil asked children to indicate whether it was "silly" or "ok" to say, for example, that "This chair is an hour long" or "This milk is alive". To separate children's judgments of matters of fact *per se* from predicability, Keil asked for judgments about individual statements and about their negations. If the child accepted either statement as "ok", Keil interpreted this as evidence that the child felt that the kind of property in question could be predicated of the thing in question. Based on children's judgments, Keil constructed what he called *predicability trees* for individual children. Four such trees, from children in different age groups, are shown in Figure 8. As the figure illustrates, Keil found that kindergarten children tend to make only two or three distinctions. As they grew older they came to differentiate more and more finely among the different types of concepts, as indicated by the restrictions they placed on what

[3] All of the simulation results reported here were fresh runs of the Rumelhart model, carried out by us, using the **bp** program of McClelland and Rumelhart (1988). We thank Rumelhart for supplying the pattern files used in his earlier simulations. Weights were initialized with values distributed uniformly between $-.5$ and $+.5$, and were updated after every pattern presentation with no momentum. The learning rate parameter ϵ was set to 0.1. Targets for learning were $.95$ for units that should be "on" and $.05$ for units that should be "off".

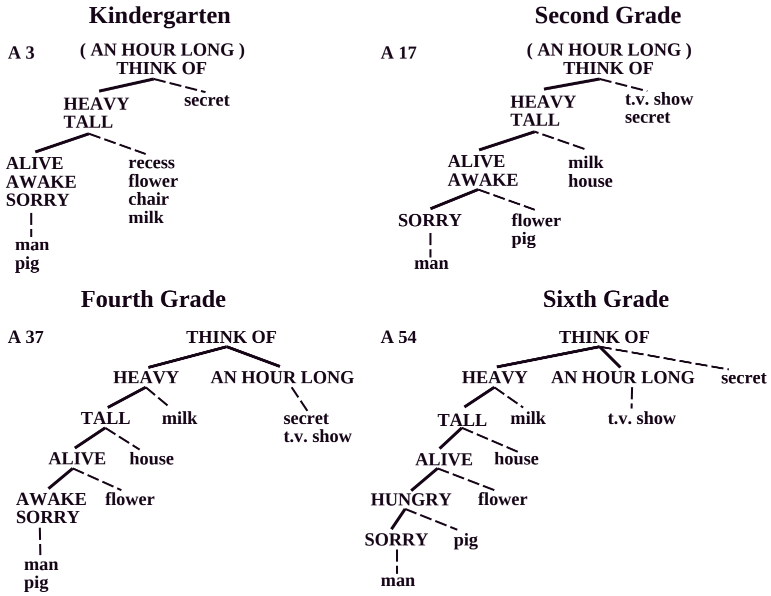

Figure 8: Examples of predicability trees empirically derived by Keil (1979). The trees indicate the types of predicates children of different ages are willing to accept as applicable to different types of concepts at different ages. The trees were derived by asking children to whether they thought statements like “This chair is an hour long” were “silly”. See text for further discussion. *Note*: Redrawn from *Semantic and Conceptual Development: An Ontological Perspective* (A3 from p. 181, A17 from p. 183, A37 from p. 185, and A54 from p. 187), by F. C. Keil, 1979, Cambridge, MA: Harvard University Press. Copyright 1979 by Harvard University Press. Permission pending.

can be predicated of what.

Keil’s (1979) developmental findings mirror the progressive differentiation of concepts that we have seen in the connectionist model. The model illustrates how conceptual distinctions can emerge as a result of very gradual training, and provides an important starting place for an experience-based approach to cognitive development. The ability to discover appropriate representations for concepts and to use them to respond appropriately to novel questions is a fundamental achievement of connectionist systems, and allows them to reopen questions about what kinds of knowledge can be acquired from experience and what must be taken to be innate (McClelland, 1994).

Catastrophic Interference

The achievements of interleaved learning systems that we have just reviewed do not mean that such systems are appropriate for all forms of learning. Indeed, it appears that they are not at all appropriate for the rapid acquisition of arbitrary associations between inputs and responses as is required, for example, in paired-associate learning experiments (e.g., Barnes & Underwood, 1959). When used in such tasks, connectionist systems like the one considered above exhibit a phenomenon McCloskey and Cohen (1989) termed *catastrophic interference*. Essentially the same point was also made independently by Ratcliff (1990).

To illustrate catastrophic interference, McCloskey and Cohen used a connectionist network slightly simpler than the one used by Rumelhart (1990). They were particularly interested in a paradigm called the $AB - AC$ paradigm, which is commonly used to study retroactive interference of one set of associations (AC) on recall of a set of associations previously acquired (AB). Here AB stands for a list of stimulus-response pairs of words, such as *Locomotive-Dishtowel*, *Table-Street*, *Carpet-Idea*,... and AC stands for a second such list, involving the same stimulus words now paired with different responses, such as *Locomotive-Banana*, *Table-Basket*, *Carpet-Pencil*,.... In such experiments, subjects are repeatedly exposed to all the items in a particular list. On each trial, they receive one A item and the task is to produce the corresponding item on the list currently under study; the correct answer is given as feedback after each recall attempt. This is repeated for the AB list until performance reaches a strict criterion, and then the subject is switched to the AC list. At different points in the series of exposures to the AC list, the subject is asked to try to recall the B members of each pair, thereby providing an opportunity to examine the extent of interference of AC learning on recovery of the AB associations.

McCloskey and Cohen’s network provided for a two-part input, as in Rumelhart’s network (Figure 9). One subset of the input units was reserved for representing each A term, and a second subset was used to represent what is called the list

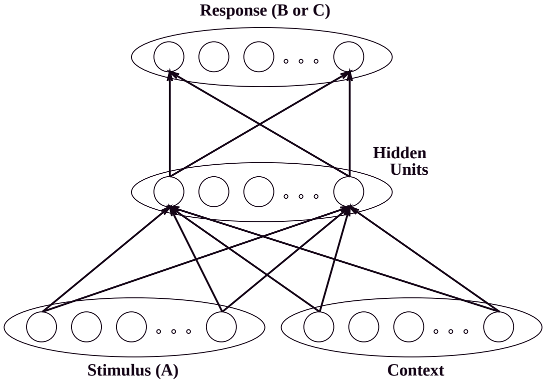

Figure 9: The network used by McCloskey and Cohen (1989) to demonstrate catastrophic interference in back-propagation networks. All output units receive connections from all hidden units and all hidden units receive inputs from both sets of input units. Based on Figure 8, page 126, of McCloskey and Cohen (1989).

context—essentially an arbitrary pattern indicating whether the items to be recalled are the B items or the C items. As in the experiment, they trained a network first on the AB list, and then shifted to AC training, testing AB performance at different points along the way. The results are shown in Figure 10a, and contrasted with typical human results in Figure 10b. The pattern McCloskey and Cohen termed catastrophic interference is evident in the network's performance. Whereas humans show a gradual loss of ability to retrieve the AB list, and are still capable of operating at over 50% correct recall after the AC list performance has reached asymptote, the network shows virtually complete abolition of AB list performance before AC performance rises above 0% correct.

One possible response to this state of affairs might be to try to find ways of avoiding catastrophic interference in multilayer networks. In fact, several investigators have demonstrated ways of reducing the magnitude of interference in tasks like those studied by McCloskey and Cohen (Hetherington & Seidenberg, 1989; Kortge, 1993; French, 1991, 1992; Sloman & Rumelhart, 1992; McRae & Hetherington, 1993). Many of these proposals amount to finding ways of reducing overlap of the patterns that are to be associated with appropriate responses via connection weight adjustment. One might then be tempted to suggest that McCloskey and Cohen simply used the wrong kind of representation, and that the problem could be eliminated by using sparser patterns of activation with less overlap. However, as French (1991) has noted, reducing overlap avoids catastrophic interference at the cost of a dramatic reduction in the exploitation of shared structure. In connectionist systems, what one learns about something is stored in the connection weights among the units activated in representing it. That knowledge can only be shared or generalized to other related things if the patterns that represent these other things overlap (Hinton, McClelland, & Rumelhart, 1986).

One could pursue the matter further, looking for ways of preserving as much of the ability to extract shared structure as possible while at the same time minimizing the problem of catastrophic interference. However, the existence of hippocampal amnesia, together with the sketch given above of the possible role of the hippocampal system in learning and memory, suggests instead that we might use the success of Rumelhart's simulation, together with the failure of McCloskey and Cohen's, as the basis for understanding why we have a separate learning system in the hippocampus and why knowledge originally stored in this system is incorporated in the neocortex only gradually.

Incorporating New Material into a Structured System of Knowledge through Interleaved Learning

To begin to address this issue, let us consider the incorporation of new knowledge into a structured system. McCloskey and Cohen's simulation does not relate to structured knowledge, since the associations being learned are arbitrary paired associates, arbitrarily grouped into lists. this issue can be explored, however, in the context of the semantic network simulation. We will see that attempts to acquire new knowledge all at once can lead to strong interference with aspects of what is already known. But we shall also see that this interference can be dramatically reduced if new information is added gradually, interleaved with ongoing exposure to other examples from the same domain of knowledge.

We illustrate these points by examining what happens if we teach Rumelhart's network some new facts that are inconsistent with the existing knowledge in the system: The facts in question are that penguins are birds, but they can swim and cannot fly. We will consider two cases. The first one we will call *focused learning*, in which the new knowledge is presented to the system repeatedly, without interleaving it with continued exposure to the rest of the database about plants and animals. We compare this to *interleaved learning*, in which the new information about penguins is simply added to the training set, so that it is interleaved with continued exposure to the full database. We use the same learning rate parameter in the two cases. We see that with focused learning, the network learns the material about penguins much more rapidly than in the case of interleaved learning (Figure 11a). In this graph, we use a measure called the absolute error, which reflects the mismatch between the network's output and the correct response. The absolute error is the sum, across all output units, of the absolute value of the difference between the correct and obtained activation. The axis is inverted so that the upward direction represents better performance, and it is apparent that learning proceeds more rapidly in the focused case. However, as we teach the network this new information, we can continue to test it on the knowledge it had previously acquired about other concepts.

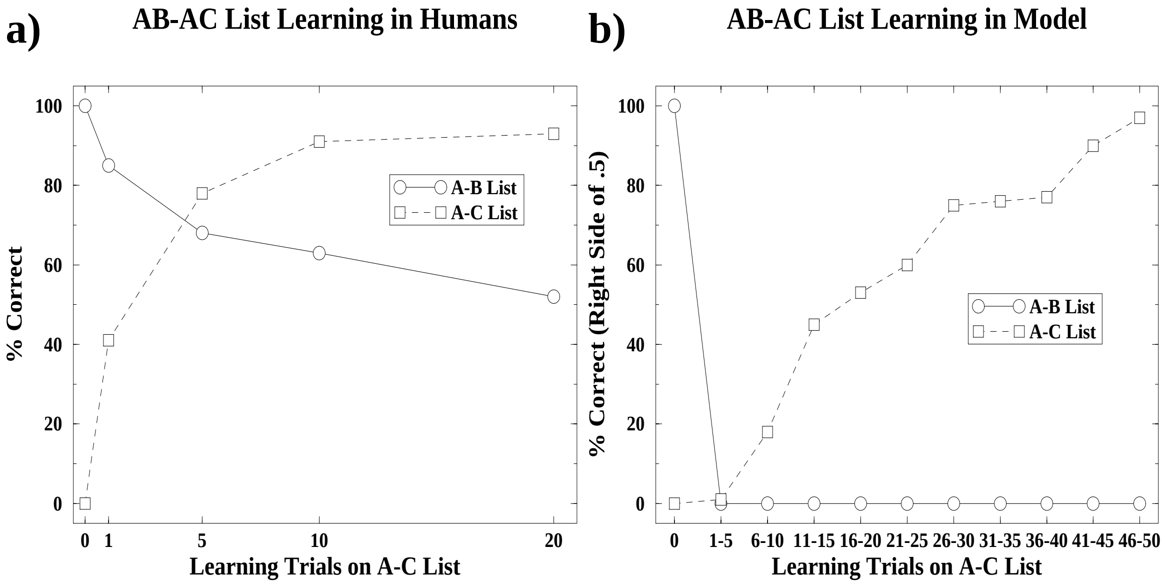

Figure 10: a) Experimental data showing mild interference in humans in the AB-AC paradigm (Barnes & Underwood, 1959), and b) simulation results demonstrating catastrophic interference. *Note*: From Figures 7 and 10a of "Catastrophic interference in connectionist networks: The sequential learning problem" (pp. 125 and 129), by M. McCloskey and N. J. Cohen, in *The Psychology of Learning and Motivation*, edited by G. H. Bower, 1989, New York: Academic Press. Copyright 1989 by Academic Press. Permission pending.

What we see is a deleterious effect of the new learning on the network's performance with other concepts (Figure 11b). The measure in the case is the average absolute error over all of the cases in which any concept (including a subordinate concept such as *robin* or *pine* or a superordinate concept such as *bird* or *animal*) is paired with the relation *can*. What happens is that as the network learns that the penguin is a bird that can swim but not fly, it comes to treat all animals—and to a lesser extent all plants—as having these same characteristics. In the case of the fish, the effect is actually to improve performance slightly, since the penguin can do all of the same things the fish can do. In the case of the birds, of course, the effect is to worsen performance a great deal, specifically on the output units that differ between the birds and the penguin. The interference is not quite as catastrophic as in the McCloskey-Cohen simulation, but it is far greater than what we see with interleaved learning.

With interleaved learning, incorporation of knowledge that penguins can swim but not fly is very gradual, in two ways. First the process is extended simply because of the interleaving with continued exposure to the rest of the corpus; second, the rate of progress per exposure, as shown in the figure, is slowed down; However, this procedure has a great benefit. It results in very little interference. Eventually, with enough practice, the network can in fact learn to activate strongly the correct output for the input *penguin-can*, and it learns to do so without ever producing more than a slight hint of interference with what it already knows about other concepts. This is because the interleaved learning allows the network to carve out a place for the penguin, adjusting its representation of other similar concepts and adjusting its connection weights to incorporate the penguin into its structured knowledge system.

We will argue below that these effects are not just idiosyncratic characteristics of back-propagation networks, but apply broadly to systems that learn by adjusting connection weights based on experience. Dramatic confirmation of catastrophic effects of focused learning in real brains—and of the benefits of interleaved learning—can be found in recent work of Merzenich (personal communication, January, 1995). He has found that highly repetitive sensory-motor tasks corresponding to focused learning lead to severe loss of differentiation of the relevant regions of sensory cortex: Practice produces a dramatic reduction in the diversity of responses of the neurons in these regions. This loss of differentiation was accompanied by a clinical syndrome called *focal dystonia*, which is a breakdown of sensory-motor coordination of the affected limb. This syndrome can be corrected in both monkeys and humans by physical therapy regimens that involve interleaved practice on a battery of different exercises.

The observation that interleaved learning allows new knowledge to be gradually incorporated into a structured system lies at the heart of our proposals concerning the role of the hippocampus in learning and memory. We see this gradual incorporation process as reflecting what goes on in the neocortex during consolidation. This view is quite close to the view of the consolidation process as it was envisioned by Squire, Cohen,

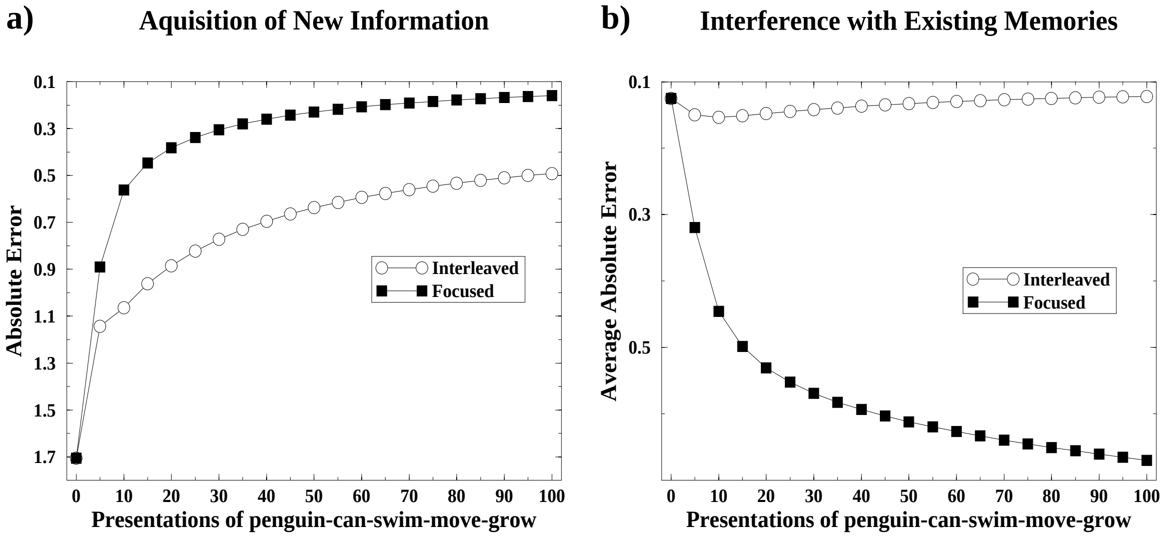

Figure 11: Effects of focused and interleaved learning on the acquisition of new knowledge and on interference with existing knowledge. Simulations were carried out using Rumelhart's network, using the connection weights resulting from the initial 500 epochs of training with the base corpus. The performance measure, absolute error, is defined as the sum across output units of the absolute value of the difference between the correct response for each pattern and the actual response. The measure reaches its optimal value of 0 when the output exactly matches the target. Better performance corresponds to lower error, and the axis is inverted for better visual correspondence to standard memory performance curves. In the analysis of interference with other memories, the performance measure is the average of the absolute error over all 15 of the cases in the initial training corpus involving the *can* relation. The scales of each graph are different, and are set to encompass the range of values spanned in each case. The interference is much greater for some items than for others, and falls predominantly on those output units where the correct answer for the pre-existing memory differs from the correct answer for the penguin.

and Nadel (1984):

> …it would be simplistic to suggest that any simple biological change is responsible for consolidation lasting as long as several years, as indicated by the data from retrograde amnesia. Rather, this time period, during which the medial temporal region maintains its importance, is filled with external events (such as repetition and activities related to original learning) and internal processes (such as rehearsal and reconstruction). These influence the fate of as-yet unconsolidated information through remodeling the neural circuitry underlying the original representation. (p. 205)

Three Principles of Connectionist Learning

The simulations presented above suggest three principles of learning in connectionist systems:

- The discovery of a set of connection weights that captures the structure of a domain and places specific facts within that structure occurs from a gradual, interleaved learning process.
- Attempts to learn new information rapidly in a network that has previously learned a subset of some domain leads to catastrophic interference.
- Incorporation of new material without interference can occur if new material is incorporated gradually, interleaved with ongoing exposure to examples of the domain embodying the content already learned.

Answers to the Key Questions

These principles allow us to formulate answers to the key questions about the organization of memory raised above:

- Why do we need a hippocampal system, if ultimately performance in all sorts of memory tasks depends on changes in connections within the neocortical system? Why are the changes not made directly in the neocortical system in the first place?

The principles indicate that the hippocampus is there to provide a medium for the initial storage of memories in a form that avoids interference with the knowledge already acquired in the neocortical system.

- Why does incorporation of new material into the neocortical system take such a long time? Why are the changes to neocortical connections not made more rapidly, shortly after initial storage in the hippocampal system?

Incorporation takes a long time to allow new knowledge to be interleaved with ongoing exposure to exemplars of the existing knowledge structure, so that eventually the new knowledge may be incorporated into the structured system already contained in the neocortex. If the changes were made rapidly, they would interfere with the system of structured knowledge built up from prior experience with other related material.

Generality of the Relation between Discovery of Shared Structure and Gradual, Interleaved Learning

Thus far we have used a very specific example to consider discovery of shared structure through interleaved learning and catastrophic interference in focused learning. We selected this example to provide a concrete context in which to make these points and to illustrate as clearly as possible how much is at stake: Our claim is that experience can give rise to the gradual discovery of structure through interleaved learning but not through focused learning and that this gradual discovery process lies at the heart of cognitive, linguistic, and perceptual development.

In this section, we examine the issues more generally. First we consider what it means to discover the structure present in a set of inputs and experiences. Then we consider general reasons why the extraction of structure present in an ensemble of events or experiences requires slow learning. To conclude the section, we discuss the process of discovering structure in biologically realistic systems.

What is structure?

Throughout this article we discuss the structure present in ensembles of events. What we mean by the term *structure* is any systematic relationship that exists within or between the events which, if discovered, could then serve as a basis for efficient representation of novel events and/or for appropriate responses to novel inputs. Marr (1970) noted that events almost never repeat themselves exactly, but yet we do learn from past experience to respond appropriately to new experiences. If there is no structure—no systematicity in the relationship between inputs and appropriate responses—then of course there will be no basis for responding appropriately to novel inputs. But if a systematic relationship does exist between inputs and appropriate responses, and if the organism has discovered that relationship, then appropriate responding may be possible.

We can begin to make this point explicit by continuing within the domain of concepts about living things. In the Rumelhart model, the structure is the set of constraints that exist on the correct completions of propositions, given a concept and a relation term. For example, if something is a bird, then it has wings and it can fly. In a symbolic framework, such constraints are captured by storing propositions that apply to entire subtrees just once at the top of the subtree; similarity relations among concepts are captured by placing then in neighboring locations in the tree. In the connectionist framework, such constraints are captured in the connection weights, and the similarity relations among concepts are captured by using the weights to assign similar distributed representations. The patterns involving the concept *sparrow* conform, by and large, to the constraints embodied in the patterns involving the concepts for *robin* and *canary*, and therefore, once *sparrow* is assigned a representation similar to the representations of *robin* and *canary*, appropriate representation and completion of propositions involving *sparrow* are possible.

In other domains, different kinds of structure can be found. For example, the English spelling system provides a notation that has a quasi-structured relation to the sound of English words. Once one has learned this structure from examples of existing words (including, for example, *save*, *wave*, *cave*, *slave*, etc.) one can generalize correctly to novel forms (such as *mave*). As a third example of structure, consider redundancies present in visual patterns. Neighboring points tend to have similar depth, orientation, and reflectance properties; such sets of neighboring points define surfaces of objects. Similarly, if there is a discontinuity between two adjacent points in the pattern, the same discontinuity will tend to exist between other pairs of adjacent points close by; such sets of neighboring discontinuities define edges. The surfaces and edges constitute structure, and given that the objects in images contain surfaces bordered by edges, it is efficient to represent images in terms of the properties and locations of the surfaces and edges. Such representations can be very efficient and can allow for completion of occluded portions of novel visual patterns.

Finally, an abstract but general example of structure is any correlation that may exist between particular pairs or larger sets of elements in a set of patterns. Such correlations, if discovered, could then be employed to infer the value of one member of the set of elements from the values of the other members, when a novel but incomplete pattern is presented. Furthermore, the presence of these correlations means that the patterns are partially redundant. This in turn means that we can represent patterns that exhibit these correlations by storing a single value for each correlated set of elements, rather than the elements themselves, as is done in principal components analysis.

Why Discovering Structure Depends on Slow Learning

Now that we have defined what we mean by structure, we are in a position to consider general reasons why the discovery of structure depends on gradual, interleaved learning. The reasons we will consider are largely independent of specifics of the network organization, the training environment, or even of the learning algorithm used.

The first reason applies generally to procedures with the following characteristics:

- The procedure is applied to a sequence of experiences, each representing a sample from an environment that can be thought of as a distribution or population of possible experiences.
- The goal of learning is to derive a parameterized characterization of the environment that generated the sequence of samples, rather than to store the samples themselves.
- What is stored as a result of applying the procedure is not the examples, but only the parameterized characterization. As each new example is experienced, the parameterized characterization is adjusted, and that is the only residue of the example.
- The adjustment process consists of a procedure that improves some measure of the adequacy of the parameterized characterization, as estimated from the data provided by the current training case.

We might call procedures with these characteristics stochastic, on-line, parameter updating procedures, but we will call them simply *stochastic learning procedures* to emphasize their relation to the question of learning and memory. In such procedures, we shall see that gradual learning is important if the parameterized characterization is to accurately capture the structure of the population of possible training examples.

Our analysis of this issue derives from an analysis of connectionist learning procedures due to White (1989). In White's analysis, the array of connection weights stored in a network is viewed as a multi-valued parameterized characterization, or *statistic*, thought to be an estimate of the weights appropriate for the entire environment or population from which actual experiences or training examples are drawn. In the case of the gradient descent learning procedure used in the semantic network model, the statistic is an array of connection weights $\mathbf{w}$ that is construed to be an estimate of the array of weights $\mathbf{w}*$ that minimizes the error measure over the population of input-output patterns. When there is more than one array of weights equivalently good at minimizing the error measure, then $\mathbf{w}$ is construed as an estimate of some member of the set of such equivalent arrays. To find an estimate that exactly matches one of these arrays of weights would be to capture all of the structure, not of the training examples themselves, but of the entire distribution from which they were drawn.

There are of course a very large number of different connectionist learning rules that can be viewed as a method for computing some statistic from the sample of training experiences. These statistics can in turn be viewed as representing some aspect of the structure present in the training experiences. Let us consider for example a version of the Hebbian learning rule that computes an estimate of the *covariance*, the average value of the product of the activations of the units on the two sides of the connection weight to unit i from unit j. The covariance is a statistic that captures one aspect of the structure present in the patterns from which it was computed. The learning rule for estimating the covariance is:

$$\Delta w_{ij} = \epsilon(a_i a_j - w_{ij}) \qquad (1)$$

Here w_{ij} is the weight to unit i from unit j, and Δw_{ij} represents the change in this weight. The variables a_i and a_j represent the activations of units i and j, and ϵ is the learning rate parameter. In the case where each event consists of a sample vector $\mathbf{a}$ of activations, then the vector of weights $\mathbf{w}$ will be an estimate of the population covariance array $\mathbf{c}$, where the elements of $\mathbf{c}$ are the covariances of activations of units i and j. In this case it is easy to see how the learning rule is acting to reduce the difference between the parameterized estimate of the covariance of each pair of units (w_{ij}) and the current data relevant to this estimate ($a_i a_j$).

This covariance learning rule provides a concrete context in which to illustrate a general point: the smaller the learning rate, the more accurate the estimate will eventually be of the population value of the statistic the learning rule is estimating, in this case the population value of the covariance of a_i and a_j.

Let us suppose, in keeping with our assumptions: 1) that we want each w_{ij} to approximate the true population value of the covariance c_{ij}, and 2) that in fact the environment is a probabilistic environment so that the value of the product $a_i a_j$ varies from sample to sample. In this case, it should be obvious that the accuracy with which the connection weight corresponds to the actual population value of the covariance will vary with the size of our learning rate parameter ϵ. The only meaningful values of ϵ are positive real numbers ≤ 1. When ϵ is equal to 1, we find that each new experience totally resets the value of w_{ij} to reflect just the current experience. With smaller values, w_{ij} depends instead on the running average of the current and previous experiences. The smaller ϵ, the larger the sample of history that is the basis for w_{ij}, and the more accurate w_{ij} will eventually be as a representation of the true population value of the statistic.

The argument just given applies very generally; it is independent of the exact nature of the statistic being estimated. There are some mathematical constraints, but these are relatively technical and we refer the reader to White (1989) for further discussion. Basically, the argument depends on the fact that when each experience represents but a single, stochastic sample from the population, it is necessary to aggregate over many samples to get a decent estimate of the population statistic. Accuracy of measurement will increase with sample size, and smaller learning rates increase the effective sample size by basically causing the network to take a running average over a larger number of recent examples.

The second reason why slow learning is necessary applies to cases with an additional characteristic beyond those listed above:

- The procedure adjusts each parameter in proportion to an estimate of the derivative of the performance measure with respect to that parameter, given the existing values of all of the parameters.

Such procedures can be called *gradient descent procedures*. The standard back-propagation learning procedure and the more biologically plausible procedures we will consider in the next section are procedures of this sort. Such procedures are guaranteed to lead to an improvement, but *only if infinitesimally small adjustments* are made to the connection weights at each step. The reason for this is that as one connection weight changes, it can alter the effect that changes to other connection weights—or even further changes to the same connection weight—will have on the error. This problem is especially severe in multi-layer networks, where the effect of changing a weight from an input unit to a hidden unit depends critically on the weights going forward from the hidden unit toward the output. This is one of the reasons why multi-layer networks trained with such a procedure require many passes through the whole set of patterns, even in cases where the network is exposed to the full set of patterns that make up the environment before each change in the weights. After each pass through the training set, the weights can be changed only a little; otherwise changes to some weights will undermine the effects of changes to the others, and the weights will tend to oscillate back and forth. With small changes, on the other hand, the network progresses a little after each pass through the training corpus. After each weight adjustment, the patterns are all presented again, and the best way to change each weight is re-computed, thereby assuring that progress will also be made at the next step. It should be noted that progress may be possible, even if there is some overshoot on each weight adjustment step. In such cases the actual rate of progress becomes decoupled from the size of the learning rate parameter. Given this it is important to distinguish between the value of the learning rate parameter and the effective rate of progress that results from the value chosen.

In multi-layer networks trained by a stochastic gradient descent learning procedure, both of the factors discussed here play a role. We can view the very small changes made after each pattern presentation as adding up, over many patterns, to an estimate of the best overall direction of change based both on the characteristics of the population as estimated from the sample and on the current values of the connection weights. It is necessary to make small changes, both to base the overall direction of change on stable estimates of the population statistics at each point and to avoid overshoot that can arise when changes that are too large are made. While we know of no analyses considering the circumstances that cause one or the other factor to dominate, it is clear that there are at least two reasons why the discovery of structure requires the use of a small learning rate.

Arbitrary Associations, Quasi-Regularity, and Memory for Facts and Experiences

It should be noted that connectionist networks that are capable of extracting shared structure can also learn ensembles of arbitrary associations. In cases of totally arbitrary associations connectionist models show strong advantages for interleaved over sequential learning (McCloskey & Cohen, 1989). This fact accords with the well-known advantages of spaced over massed practice of arbitrary material. In humans, massed practice tends to allow for relatively rapid initial acquisition of each association compared to the interleaved case, but this initial advantage gives way to a strong disadvantage when performance on an entire series of associations is tested on a delayed test (see Schmidt & Bjork, 1992, for a brief summary of some of the relevant evidence).

The reasons why learning ensembles of arbitrary associations requires interleaved learning are similar to the reasons why the extraction of shared structure requires interleaved learning: In both cases, the goal of the learning procedure is to find a set of connections that handles an entire ensemble of events and experiences, rather than just each individual case. With interleaved learning the direction of weight changes is governed by the entire ensemble, not just the most recent individual case, and so the outcome is successful performance on an entire ensemble of cases McCloskey and Cohen (1989) themselves made this point for the case of ensembles of arbitrary associations.

As we shall see below, the need for interleaved learning can be eliminated by exploiting totally non-overlapping representations of each example. One trouble with this scheme is that it is extremely inefficient for large systems of arbitrary associations (such as the set of largely arbitrary associations of words and their meanings). Hinton et al. (1986) showed that in such cases much greater efficiency can be achieved using overlapping distributed representations. Learning these representations, however, requires interleaved learning, and proceeds very slowly due to the lack of shared structure.

Another difficulty with using completely non-overlapping representations is the fact that total arbitrariness is the exception rather than the rule in cognitively interesting domains, and non-overlapping representations prevent the exploitation of the systematic aspects of these relationships. In general arbitrary aspects of particular associations coexist with partial regularities. For example, consider the problem of learning exception words in the same system that learns typical spelling to sound correspondences. This is a domain that can be called *quasi-regular*: it contains many items that are partially arbitrary, in that they violate some aspects of the shared structure of the domain, but not all. As an example, consider the word PINT. First of all, both its spelling and its sound consist of familiar elements. Second, in this word, the letters P, N, and T all have their usual correspondences, while the I has an exceptional correspondence. While some have argued that such items should be stored totally separately from the system of structured spelling-sound correspondences, incorporation of PINT into a structured system would allow for partial exploitation of the structure. It has now been shown that such items can be incorporated in such systems, without preventing them from handling novel items (e.g., VINT) in accordance with the regular correspondences of all of the letters, to an extent indistinguishable from English speaking college students (Plaut, McClelland, Seidenberg, & Patterson, in press).

We believe that the domains encompassed by semantic, episodic, and encyclopedic knowledge are all quasi-regular, and we suggest that facts and experiences are only partially arbitrary, similar to exception words. Consider for example John F. Kennedy's assassination. There were several arbitrary aspects, such as the date and time of the event. But our understanding of what happened depends also on general knowledge of presidents, motorcades, rifles, spies, etc. Our understanding of these things informs—indeed, pervades—our memory of Kennedy's assassination. Perhaps even more importantly, though, our understanding of other similar events is ultimately influenced by what we learn about Kennedy's assassination. It is the integration of the contents of ensembles of such experiences into structured knowledge systems that provides the substance of semantic, episodic, and encyclopedic memory.

To consolidate the contents of a partially arbitrary episode or event, the neocortical system will need to find a set of connection weights that accommodate both the common and the idiosyncratic aspects. Those aspects that are shared with other events and experiences will be the most easily consolidated—indeed, the system of connection weights may already incorporate these aspects when the association is first encountered. Those that are idiosyncratic will take more time to acquire, as is well documented in simulation studies of interleaved learning in quasi-structured domains (Plaut et al., in press). Decay of hippocampal traces over time comes to play a crucial role in this context. If the rate of decay is relatively rapid, compared to the rate of consolidation, much of the idiosyncratic content of individual episodes and events may not be consolidated at all. This race between hippocampal decay and interleaved learning thus provides the mechanism that leads to what Squire et al. (1984) describe as the schematic quality of long-term memory: arbitrary and idiosyncratic material tends to be lost, while that which is common to many episodes and experiences tends to remain. However, we should note that there is nothing preventing the consolidation of some totally arbitrary material encountered in experience only once, if it is reinstated in the neocortical system frequently enough.

Discovery of Structure in Biologically Realistic Systems

Let us now consider the process of discovering structure as it might occur in the mammalian neocortex. First, some of the structure present in ensembles of inputs can be extracted using very simple learning rules, similar to the covariance rule described above. One example of such structure is the pattern of intercorrelations among the various inputs to a neuron or group of neurons. Several researchers have proposed that the discovery of the relative magnitudes of these correlations may play a central role in the development of receptive fields and the organization of these fields into columns (Linsker, 1986a, 1986b, 1986c; Miller, Keller, & Stryker, 1989; Miller & Stryker, 1990; Kohonen, 1984). For example, Linsker (1986a) uses the following learning rule in his model of the development of center-surround receptive fields:

$$\Delta w_{ij} = \epsilon({a_i}^L - b^L)({a_j}^{L-1} - b^{L-1}) + \kappa \qquad (2)$$

In this equation, ${a_i}^L$ and ${a_j}^{L-1}$ refer to activations of two neurons in layers L and $L-1$ of a multi-layered, feedforward network, and b^L, b^{L-1}, and κ are constants that regulate the weight changes. The rule is similar to the covariance learning rule already discussed. Weights between units that are correlated more than a certain amount are increased, and other weights are decreased. Individual weights are bounded in Linsker's models, so they tend to increase over time to the upper bound or decrease to the lower bound.

The development of center-surround organization in this model occurs by assigning positive connection weights to inputs that are maximally correlated with other inputs to the same neuron. The set of inputs that are most correlated with each other come to have strong positive connections to the receiving unit, while positive connections from other input units drop away. The model depends on slow learning because otherwise many of the correlations that need to be detected would be lost in noise. The weights must change slowly enough so that their overall direction of change is governed by the true correlations. Linsker considers one case where the correlations are so small relative to the noise that it is necessary to sample about 8,000 input patterns to determine the correct direction of weight changes.

Correlations among inputs can be detected with simple local learning rules, but these rules are not necessarily adequate to learn all aspects of the structure that may be present in an ensemble of events, particularly when part of the structure lies in relations between inputs and desired outputs, which can be construed as inputs in another modality or inputs at a later point in time. Sometimes, the structure is hidden, in the sense that it is not present as a direct relationship between actual inputs and desired outputs, but only as a relationship between inputs once they have been appropriately re-represented. This situation arises, for example, in the Rumelhart (1990) semantic network model discussed above. In general, the problem is that the choice of an appropriate representation for one part of an input depends on the use to which that representation is to be put by the rest of the system. This information is simply not available within the different parts of the input considered separately, and requires some form of bi-directional communication among the different parts of the system.

The major breakthrough in connectionist learning was the discovery of procedures, more powerful than simple correlational learning rules, that could learn to form these representations (Rumelhart, Hinton, & Williams, 1986). The purpose of the procedure is to make available, at each connection in the network, information about the extent to which the adjustment of that connection will reduce the discrepancy between the actual output of the network and the desired output—i.e., the partial derivative of the error with respect to each connection weight. Each connection weight is then adjusted by this amount, and gradually—as we have seen in the semantic network example—the structure underlying the entire ensemble of patterns in the training set is discovered. As important as this learning rule has been computationally, however, there remains a road block to a synthesis of computational and neural science, since the actual procedure used to calculate the relevant derivatives seems biologically unrealistic. Rumelhart's semantic network model exemplifies the situation. Activation signals propagate in one direction, from input to output, and the process of determining the appropriate adjustments to the crucial weights from the concept input units to the concept representation units depends on a computation that appears to correspond to a biologically implausible backward transmission of a separate error signal across forward-going synapses. Because of this, the learning algorithm is tolerated in neuroscience circles as a method for finding optimal connection weights that perform some task, but it is specifically disavowed as a possible mechanism for learning in real biological systems (e.g., Zipser & Andersen, 1988). This leaves us, though, without a biologically plausible mechanism for discovering structured relations between inputs and outputs in multi-layer networks.

One solution to this problem comes from the idea that learning in multilayer systems might exploit the reciprocity of ordinary axonal projections that appears to hold between regions of the neocortex. It appears to be quite generally true that whenever there are connections from region A to region B there are also connections returning from region B to region A (Maunsell & Van Essen, 1983). Such return connections can allow levels of processing near the input to be affected by results of processing further upstream. In fact, it has been shown in a number of different cases that the necessary error derivatives can be computed from the activation signals carried by ordinary feedback connections (Barto, Sutton, & Brouwer, 1981; Ackley et al., 1985; Grossberg, 1987; Hinton & McClelland, 1988).

For example, Hinton and McClelland (1988) showed that hidden units can calculate terms equivalent to the error derivatives used in back propagation by using the difference between the activation signals returning from output units before and after the desired output is provided to the output units. This and related procedures are generally robust in the face of incomplete reciprocal connectivity, and can even operate when the return activation is mediated by interneurons (Galland & Hinton, 1991; see also Hopfield, 1982). In fact, random initial connections subject only to relatively coarse topographic constraints of the sort that appear to typify reciprocal connectivity between brain regions can be used, and the system will naturally tend to increase the degree of symmetry (Hinton, 1989). Random synaptic sprouting coupled with pruning of unused connections could further contribute to the symmetrizing effect.

A second approach is to replace back propagation of error information with a single, diffusely propagated reinforcement signal of the kind that could easily be distributed widely throughout the brain by a neuromodulatory system. Mazzoni, Andersen, and Jordan (1991) have compared an associative reinforcement learning algorithm and the back propagation algorithm as procedures for discovering representations that are useful for the transformation of visual space from retinal to head-centered coordinates and for development of simulated neurons with response properties resembling those found in area 7a. Both procedures can be used, and both discover receptive fields of the same types that are found in the brain. Interestingly, for large-scale networks, this type of reinforcement learning appears to require even more gradual learning than back-propagation (Barto & Jordan, 1987).

It is not our intention to suggest that there exists any complete understanding of the exact procedures used by the brain to discover the structure present in ensembles of patterns. Our argument is only that procedures that compute the relevant information must exist, and some such procedures have been proposed that are quite biologically plausible. Whatever the exact procedure turns out to be, it will involve slow, interleaved learning. The reason is simply that structure is not in fact detectable in individual patterns, but necessarily requires information that is only present in ensembles of patterns. Interleaved learning allows connection weight changes to be governed by this sort of information.

Combining the Hippocampal and the Neocortical Learning Systems: Consolidation and Retrograde Amnesia

We have seen how it is possible, using interleaved learning, to gradually discover the structure present in ensembles of events and experiences, and to integrate new knowledge into the connection weights in a system without producing interference with what that system already knows. The problem is that acquiring new information in this way is very slow—and if the cortical system works like the systems we have discussed, it would obviously be insufficient for meeting the demands of everyday life, in which information must often be acquired and retained on the basis of a single exposure. Our argument is that it is precisely to allow retention of the contents of specific episodes and events, while at the same time avoiding interference with the structured knowledge held in the neocortex, that the hippocampus and related structures exist. As we have already reviewed, these structures are crucial for the rapid formation of memory traces for the contents of specific episodes and events.

Once a memory is stored in the hippocampal system, it can be reactivated and then reinstated in the neocortex. Such reinstatements will have two important consequences: First, reinstatement of the stored event in appropriate contexts would allow the reinstated pattern to be used for controlling behavioral responses (e.g., uttering the name of the person in front of us, when we have previously stored that name in association with the face). Second, reinstatement provides the opportunity for an incremental adjustment of neocortical connections, thereby allowing memories initially dependent on the hippocampal system to gradually become independent of it.

Experimental studies of consolidation generally use relatively arbitrary pairings of stimuli with other stimuli and/or responses. For example, the experiment of Zola-Morgan and Squire (1990) that we will discuss below requires animals to learn totally arbitrary associations between food pellets and junk objects. It appears consolidation of such arbitrary material occurs through the same process of gradual incorporation into the neocortical structures that is used for learning more structured material. As previously discussed, consolidation of arbitrary material allows efficient representation, and even experiences that have arbitrary elements generally share some structure with many other experience that gradual consolidation will allow the system to exploit.

A key question arises as to the source of reinstatements of exemplars drawn from the existing knowledge structure, since their interleaving with new knowledge is crucial for the prevention of catastrophic interference. There are several (nonexclusive) possibilities, including direct reinstatement from the external environment and re-activation in the course of cognitive activity or reminiscence. In addition, spontaneous reactivation in the absence of external input may be possible. As discussed in an earlier section, multiple single-neuron recording in hippocampus suggests such spontaneous reactivation during slow wave sleep (Wilson & McNaughton, 1994b), and a similar process of reinstatement could apply to patterns arising from the structured knowledge in the neocortical system. Possibly, events reactivated in hippocampus during slow-wave sleep prime related neocortical patterns, so that these in turn become available for activation during REM sleep. This could permit both new and old information to be played back in closely interleaved fashion.

Modeling Temporally Graded Retrograde Amnesia

To illustrate our conception of the consolidation process, we undertake in this section to provide simulations of two experiments in the growing literature on retrograde amnesia. In both studies, the manipulation is a bilateral lesion to some or all of the hippocampal system at some time after exposure to some learning experience.

In the simulations that follow, we do not actually attempt to simulate the formation of memories in the hippocampal system in a network model. Such a simulation would have the virtue of forcing us to demonstrate the mechanistic feasibility of our account, but to be at all faithful to the complexity of the hippocampal system would require a level of detail that would tend to take us away from our current focus on the systems level. Therefore we treat the hippocampus as a 'black box' that is assumed to carry out the functions we previously ascribed to it, and we concentrate on showing how an account may be provided of much of the existing data, in terms of a relatively small number of assumptions about the storage and decay of memory traces in the hippocampal system and their reinstatement in the neocortex.

The key assumptions underlying the present simulations are the following. First, we assume that hippocampal learning is a matter of degree that depends on the salience or importance of the original episode or event. Second, we assume that, as time passes, hippocampal memory traces degrade, becoming effectively weaker with the passage of time. This could occur as a result of passive decay of the relevant enhanced connections and/or as a result of interference caused by new learning. Third, we assume that the probability of hippocampally-mediated reinstatement in the neocortex decreases with the strength of the hippocampal trace. Finally, we assume that probability of reinstatement in a given amount of time may be different in task-relevant and task-irrelevant contexts. On a moment-by-moment basis, reinstatement is assumed to be more likely in task-relevant than in task-irrelevant contexts, since probe patterns generated in the former will be more similar to the pattern stored in memory than probe patterns generated in the latter, at least on the average.

A complicating factor for modeling consolidation is the fact that reinstatement of a pattern in the hippocampal system might strengthen the hippocampal representation as well as the representation in the neocortex. This could greatly retard the decay of the hippocampal trace. In this context, however, it is of interest to note that there is evidence that hippocampal synaptic plasticity is suppressed during some phases of sleep (Leonard, McNaughton, & Barnes, 1987). This suggests the possibility that at least some spontaneous reinstatements in task-irrelevant contexts may not be self-reinforcing. If task-relevant reinstatements were self-reinforcing but spontaneous reinstatements were not, this would provide a mechanism whereby memories that remain relevant would tend to persist longer in the hippocampal system than memories of only transitory relevance. In any case, the remainder of this section ignores the effects of self-reinforcement for simplicity. Such effects, if they exist, would tend to slow the apparent rate of decay from the hippocampal system.

The modeling work described below makes use of the foregoing ideas in the following way. We use the assumptions just given to justify specific training regimes for simple neural-network analogs of the neocortical systems that we assume underlie performance of the tasks animals are asked to perform in particular experiments, and show how consolidation arises under these assumptions. The hippocampus is not implemented, but is instead treated as a source of training data for the model neocortical networks. The networks used are simple, generic three-layer networks of the kind used by McCloskey and Cohen (1989). Learning in such networks occurs through repeated presentations of patterns, interleaved with other patterns. In our simulations, hippocampus-generated presentations to the cortical network due to experimenter-determined learning experiences are assumed to be interleaved with ongoing exposure to the contents of other experiences and events. As previously noted this ongoing exposure to such background patterns probably depends on reinstatements from memory as well as direct input from the environment, but there is no data on the extent of hippocampal involvement in this process. For simplicity therefore the simulations below treat the rate of ongoing exposure to such patterns as independent of the status of the hippocampal system.

Kim and Fanselow (1992). Kim and Fanselow (1992) studied the role of the hippocampal system in memory consolidation in rats. Each animal was placed in a novel environment, where it was exposed to 15 pairings of a tone with foot-shock, and then returned to its home cage. After 1, 7, 14, or 28 days they received either bilateral hippocampal lesions or sham lesions (as another control one further group received neocortical lesions at 1 day post learning). Seven days after surgery, the animals were reintroduced to the environment in which the tone-shock pairs had been presented, and their apparent fear of the situation was monitored (percent of time spent in typical fear postures), in the absence of any presentation of either tone or shock. The data are shown in Figure 12a. There were no reliable effects of delay in the sham lesioned group, although there was a trend toward a decrease. The hippocampal animals, however, showed hardly any fear if they received a hippocampal lesion one day after the tone-shock experience. There was a clear increase in the fear response as a function of time between experience and lesion, demonstrating a consolidation process that apparently extended over the full 28-day period.

As a simulation analog of consolidation in this situation, we used a three-layer network consisting of 16 input, 16 hidden, and 16 output units, and trained it on a set of 20 random stimulus-response associations (i.e., 20 input-output pairs, each consisting of a random pattern of 1's and 0's). We took these associations to represent other experiences of the animal. We assume that the neocortical system continues to be exposed to these associations throughout. For simplicity we treat the rate of recurrence of each pattern as constant over time, with each pattern occurring exactly once per simulated day of the experiment. We then added to this training corpus one additional training pair, analogous to the environment-tone-shock association; therefore we call this the ETS pair. After the experimental exposure to this association it would be available only via the hippocampus. After introduction of the new pair, training continued as before, analogous to the exposure of the cortical system to the new pattern interleaved with continued presentation of other memory traces. Although it is one of our assumptions that hippocampal traces generally decay with time, we ignore this decay for simplicity in this initial simulation, since it appears from the lack of a significant effect of delay in the control animals that the hippocampal trace is decaying very slowly if at all in this case. Thus the hippocampal trace of the new experience remains at full strength for the duration of the experiment (in control animals) or until the hippocampus is removed (for hippocampal groups).

In this initial simulation our interest focuses on the performance of the lesioned animals, since this illustrates the consolidation process. We monitored the response of the network to each presentation of the new pair, and the performance of the network is graphed in Figure 12b. Accuracy of the network's response is measured as the reduction in the average squared deviation from the correct ETS output pattern, as a fraction of the initial deviation obtained prior to any exposure to this pattern. The figure illustrates the gradual incorporation of the new association into the simulation analog of the neocortical system. We can compare the network's progress in learning the new association with the performance of Kim and Fanselow's rats who received hippocampal lesions at different points after exposure to the ETS combination. For this comparison, we have transformed the data from experimental animals into a comparable measure of proportion of maximal response, which we assume is reflected in the mean time spent freezing averaged across the control conditions. The learning rate parameter in the simulation was adjusted to produce an approximate fit to the data with one epoch of training corresponding to one day between exposure and hippocampectomy. The simulation follows an approximately exponential approach to maximal performance that falls within the error bars of the experimental data.

The details of the frequency and timing of reinstatement are of course completely unknown. The simulation indicates that it is possible to account for Kim and Fanselow's consolidation data by assuming a constant rate of reinstatement over time, and no actual hippocampal decay in this case. Various other assumptions are also consistent with the data, however. For example, there is a slight indication of some reduction in freezing with delay in the control animals, suggesting perhaps that the hippocampal trace might have weakened to some extent with time. If so, we would expect a gradual reduction in the frequency of reinstatement, and this in turn would lead to a consolidation curve with a somewhat sharper initial rise relative to the slope of the curve over the later phases of the consolidation period (we explore this matter more fully in a subsequent section). Such a pattern is consistent with, though hardly demanded by, the data, given the size of the error bars around the points.

Zola-Morgan and Squire (1990). Zola-Morgan and Squire (1990) obtained evidence of consolidation over a period of about 10 weeks in monkeys. They trained monkeys on a set of 100 binary discriminations. Each discrimination involved a pair of junk objects, one of which was consistently reinforced and the other of which was not. Animals were trained on five successive sets of twenty of these discriminations. For each set of 20, each animal was trained on two of the discriminations on each of ten successive days. The training sessions in the first set occurred an average of 15 weeks prior to surgery; those in the other sets occurred an average of 11, 7, 3, or 1 week before surgery. At the end of each set of 20 discriminations, the animal received one more exposure to each discrimination as a final test. At the end of training, 11 animals had the hippocampus as well as entorhinal and parahippocampal cortex removed bilaterally, and seven had sham lesions. Two weeks later, all animals were tested on all 100 discriminations each presented once over two 50-trial sessions.

The experiment produced a fairly standard if somewhat noisy forgetting curve for the normal controls, with accuracy dropping from about 80% for the discriminations learned an average of 1 and 3 weeks prior to surgery to about 70% for discriminations learned 11-15 weeks prior to surgery (see Figure 13). The animals with hippocampal lesions, on the other hand, showed performance in the low sixties for the discriminations learned an average of one week prior to surgery, but this increased to a peak of about 70% at 11 weeks, indicating that there was some consolidation over about a 10 week period between initial learning and hippocampal removal. Given the lack of a difference between the lesioned animals and controls at or beyond 11 weeks, it would appear that the hippocampal contribution becomes negligible at about that point.

We simulated this experiment using a three-layer network consisting of 50 input units, 15 hidden units, and a single output unit. The network was trained on 100 input-output pairs. Each input pattern consisted of two random 25-element patterns treated as corresponding to the two junk objects in each discrimination in the Zola-Morgan and Squire (1990) experiment. The random patterns were constructed simply by setting each of the 25 elements to 1 with probability 0.2 or to 0 with probability 0.8. This makes the patterns somewhat sparse and therefore somewhat distinctive. The two random patterns were concatenated to form a 50-element input pattern. Either the first or the second object in the pair was designated correct;

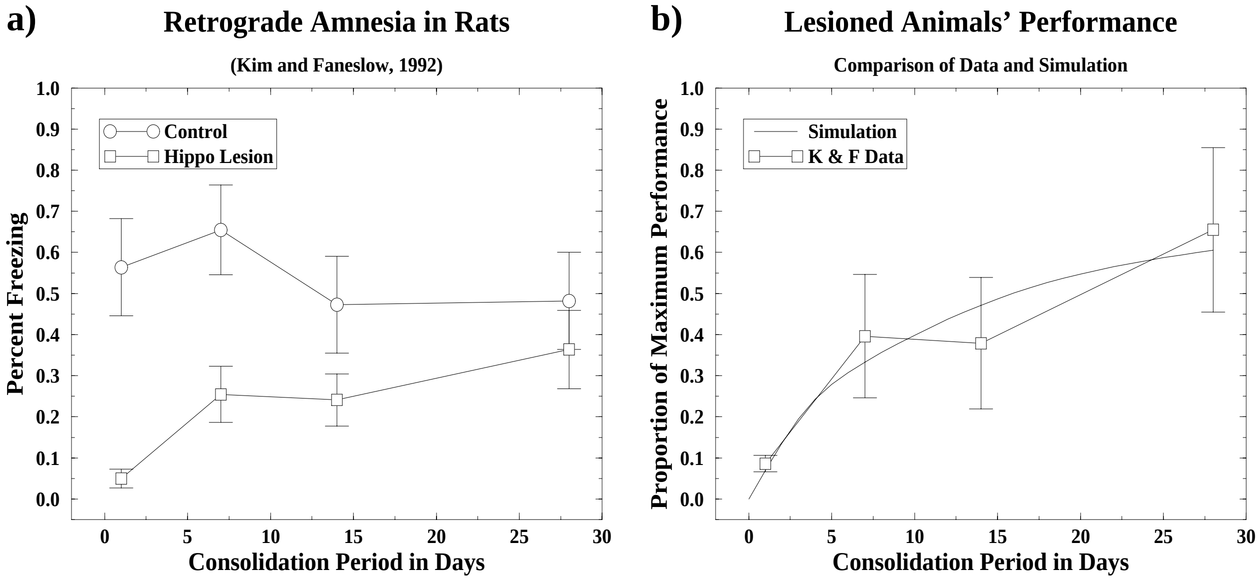

Figure 12: Consolidation in the experiment of Kim and Fanselow (1992). a) Experimental data from experimental and control groups. b) Simulation of consolidation in the group with hippocampal lesions. This panel shows the experimental data with error bars, together with a curve describing the buildup of performance over the consolidation in the simulation. For this panel, the measure used for the experimental data is the amount of time spent freezing for each experimental group, divided by the average amount of time spent freezing across the control groups at different delays. The measure used for the simulation divides the reduction in output mean squared error at each test point by the initial amount of error to the test input prior to any learning on this pattern. *Note:* Data in (a) are from Figure 2 of "Modality-specific retrograde amnesia of fear", by J. J. Kim and M. S. Fanselow, 1992, *Science, 256*, p. 676. Copyright 1992 by the American Association for the Advancement of Science. Permission pending.

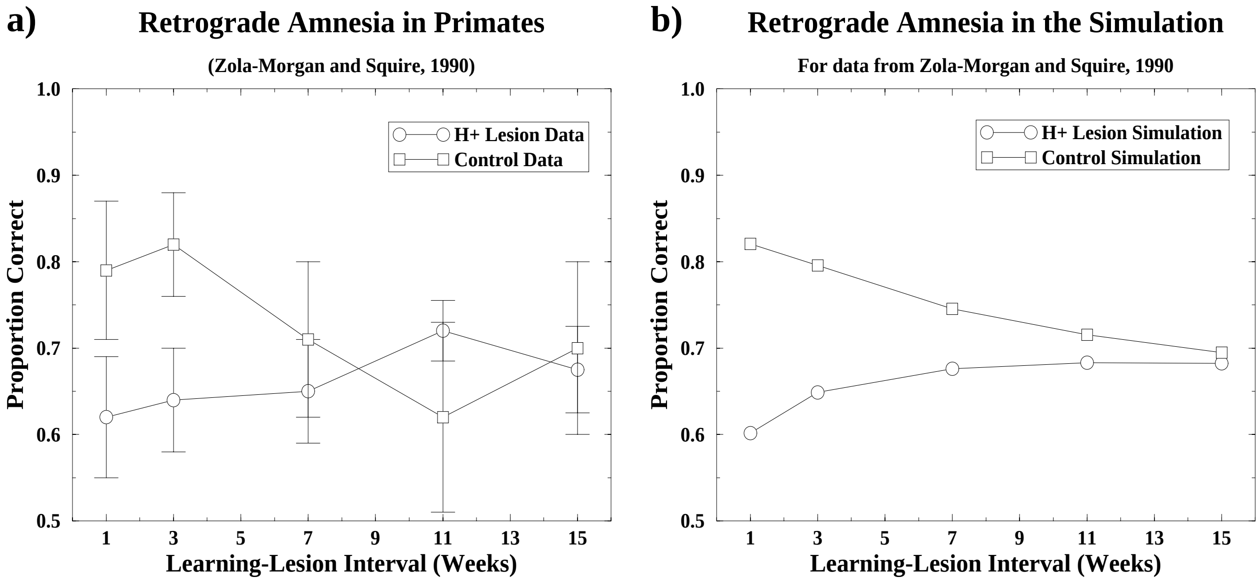

Figure 13: Consolidation in the experiment of Zola-Morgan and Squire (1990). (a) Experimental data from animals with hippocampal lesions and control animals. (b) Results of the simulation of this experiment described in the text. *Note*: Data in (a) are from Figure 2 of "The primate hippocampal formation: Evidence for a time-limited role in memory storage", by S. Zola-Morgan and L. R. Squire, 1990, *Science, 250*, p. 289. Copyright 1990 by the American Association for the Advancement of Science. Permission pending.

we trained the net to turn the output unit on if the first object is correct and off if the second object is correct (assignment of which object was correct was random with equal probability for the two objects). To test the network we simply present each input and observe the activation of the output unit. If this is greater than 0.5 the net is taken to have chosen the first object; otherwise it is taken to have chosen the second.

The training regime attempted to capture a set of training events consistent both with our theory and with the design of the experiment. As in the case of the Kim and Fanselow simulation, we presented the network with an epoch of training for each day of the experiment. Each day's training contained three types of trials: Background trials, representing ongoing exposure to a constant environment; direct-experience training trials, corresponding to the actual experiences of Zola-Morgan and Squire's animals in the training trials themselves; and reinstated experience trials, corresponding to reinstatement of experiences from the experiment via the hippocampus. The background trials began 100 simulated "days" before the experiment proper and continued for the 109 days of the experiment. There were a total of 250 background items, and each of these was sampled with a probability of 0.2 per day, so that on the average there were 50 such background items per day. The direct experience trials exactly mirrored the training regime used by Zola-Morgan and Squire, so that on the first day there were 14 presentations of the first discrimination followed by 14 presentations of the second, and so on. The reinstated-experience trials were determined as follows. For each direct-experience, a hippocampal trace was assumed to be formed. These traces were assumed to start at a nominal strength of 1 and decay at a fixed rate D per day. On each day prior to hippocampal surgery stored traces were reinstated with a probability equal to the strength of the trace times a reinstatement probability parameter r. After surgery, no further consolidation based on hippocampal traces occurred. However, the exposure to the background environment continued as before.

To model the performance of the controls, we assumed that in their case consolidation continued through the sham lesion and on for the next 14 days until testing occurred. In addition, we assumed that performance could be based on retrieval from the hippocampus. If hippocampal retrieval failed, we assumed performance would be based on the output of the neocortical network. For retrieval from the hippocampus, each of the stored traces of the same discrimination was tested for retrieval, and if any one of these was successful, retrieval was considered successful. For each trace, the probability of retrieval was equal to the strength of the trace given the time since initial study, times a retrieval probability parameter R. Note that this R reflects the probability of retrieving a trace during a test trial, given as a retrieval cue the presentation of the two relevant junk objects. It is quite different from r, the probability of reinstatement of a trace over the course of an entire 24 hour period, but in the absence of any particular cue. The only other parameters of the simulation were the hippocampal decay rate parameter D, and the cortical learning rate parameter ϵ. Due to the randomness inherent in the patterns and the training experience, there is considerable variability in the simulation. To compensate for this each simulation run involved 200 simulated subjects per condition. Several runs were performed with different values of the parameters.

The results of the best fitting simulation run are shown in Figure 13. Given the variability in the Zola-Morgan and Squire data, it is hard to tell whether the deviations between the model and the data should be taken at all seriously. From the point of view of the simulation, the data points for both groups at 11 weeks seem particularly anomalous; for both the normal and the lesioned groups they represent the largest discrepancies from the data. Since the simulated data points all fall within or near one standard error of the mean of each data point, there is no statistical basis for thinking that these anomalies are meaningful. The values of the free parameters of the simulation are instructive: though the data are noisy, sizable changes to these parameters do result in much poorer fits. First, the value of the learning rate parameter ϵ was 0.03. With this value, learning occurs very gradually indeed. The decay rate D for hippocampal traces was 0.025 per day. At this rate, hippocampal trace strength is down to 1/6 of its original value in 10 weeks. The parameter r, the probability of off-line reinstatement from the hippocampus, is 0.1 per training trial per day. Given this value, each discrimination (represented by 15 separate training trials) will be reinstated about 1.5 times a day when it is fresh, dropping to an average of 0.15 times per day at 10 weeks. Including the initial cortical exposures from the direct-experience training trials, this gave a number of cortical training trials ranging from 25 for the items presented an average of one week before surgery, to 63, for items presented an average of 15 weeks before surgery. The value of R, the probability of trace reinstatement in a test trial, was 0.07; this yields a probability of 0.6 of retrieving at least one trace of a particular discrimination just at the end of training. By the time the test occurs two weeks later, the probability of retrieving at least one trace of an item in the set studied just before (sham) surgery is 0.47. This drops to 0.19 for items studied 7 weeks before surgery (9 weeks before test) and to 0.05 for the items studied 15 weeks before surgery.

The simulation may help us understand why the evidence for consolidation is in fact somewhat weak in this experiment. The simulation shows a consolidation effect—that is, a slight increase in performance as a function of lesion delay among the lesioned groups—but it is relatively small, for two reasons. First, a considerable amount of neocortical learning actually occurs in these simulations during the period allocated for training of each batch of associations. Second, the rate of decay of traces from the hippocampus appears to be high enough to force the bulk of the consolidation to occur within the first few weeks. Given the range of training-to-lesion intervals used, and the apparent rate of hippocampal decay, the experiment provides a relatively small window on the process of consolidation.

A Simplified Quantitative Formulation of the Consolidation Process

For the purposes of facilitating further thinking and research about the time course of consolidation, we have found it useful to adopt a very abstract and simplified two-compartment model of the memory storage and consolidation process. This formulation attempts to capture the quantitative relationships seen in the simulations just described in terms of a few simple equations. [4] The formulation is depicted graphically in Figure 14.

Our formulation assumes first of all that each experienced event is stored in the hippocampus with some initial strength $S_h(0)$. This initial strength ranges between 0 and 1, and the strength at time t follows an exponential decay from this initial value:[5]

$$\Delta S_h(t) = -D_h S_h(t) \tag{3}$$

The initial strength $S_h(0)$ and the decay rate D_h may depend on the task and stimulus conditions.

When the hippocampus is off-line, reinstatements that subserve consolidation will occur with some probability $\rho(t)$ per unit time. The probability of reinstatement depends on the residual strength of the trace times the reinstatement rate parameter r_h:

$$\rho(t) = r_h S_h(t) \tag{4}$$

We assume that neocortical trace strength is incremented with each neocortical reinstatement of the trace. The amount of the increment is proportional to the learning rate parameter ϵ times the difference between the current cortical trace strength and the maximum strength of 1.0. Neocortical trace strength also decays at some rate D_c. (As with the hippocampal decay, this may be passive or may result from interference produced through the storage of other traces). Taking the probability of reinstatement into account, the change in the cortical strength at each time step is given by

$$\Delta S_c(t) = CS_h(t)(1 - S_c(t)) - D_c S_c(t) \tag{5}$$

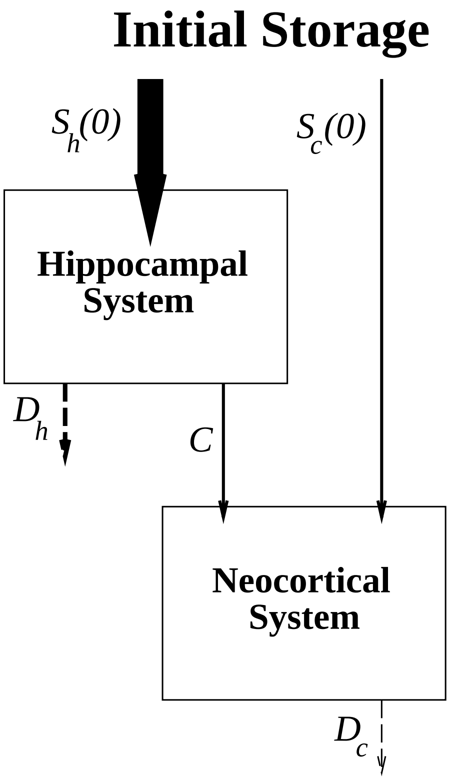

Figure 14: A simple two-compartment model that characterizes memory storage and decay in the hippocampal system, together with consolidation and subsequent decay in the neocortical system. Arrows are labeled with the parameters of the simple model: $S_h(0)$ and $S_c(0)$ refer to the strength of the hippocampal and neocortical traces due to the initial exposure to the event, D_h and D_c refer to the rate of decay from the hippocampal system and the neocortical system respectively, and C refers to the rate of consolidation.

[4] In part of a Ph. D. dissertation undertaken at about the same time as our work, Lynn (1994) also developed a simple conceptual model of the consolidation process very similar to the one presented here. He fit his model to findings from six consolidation studies, including three of the four studies considered here, and found parameter estimates broadly consistent with ours.

[5] A slightly more complex formulation would allow strengths to vary over the positive real numbers without upper bound. In this formulation some sort of non-linear function is needed to determine the probability of reinstatement of a trace given its strength. For the present the data do not provide a compelling case for introducing this complication so we have left it out for simplicity. However, it may be worth noting that this modification has the effect of reducing at least initially the effective rate of decay of hippocampal traces; the effective strength, as measured in terms of the rate of reinstatement, drops slowly at first and then drops more rapidly later, once the underlying trace strength drops into the linear range of the non-linear function.

here C is the consolidation rate, equal to the product of ϵ and r_h.

When the system is probed in some particular task context, the probability that the hippocampal trace will be reinstated in a form sufficient to drive correct, task-appropriate behavior is assumed to be given by

$$b_h(t) = R_h S_h(t) \tag{6}$$

In this equation, R_h reflects the adequacy of the task/context situation as a cue to the hippocampal memory trace. The probability that the consolidated cortical trace will be sufficient for correct task appropriate behavior is assumed to be

$$b_c(t) = R_c S_c(t) \tag{7}$$

where R_c reflects the adequacy of the task/context situation as a retrieval cue for the neocortical trace.

Correct behavior can be based on the hippocampal system, if it produces an output, or on the neocortical representation if the hippocampal system is either unavailable or does not produce a response in this case. Given this, the probability $b_{hc}(t)$ of correct behavior based on either the hippocampal or the neocortical system will be

$$b_{hc}(t) = b_h(t) + (1 - b_h(t))b_c(t) \tag{8}$$

Correct behavioral responses can also arise due to pre-existing tendencies, or to random choice when faced with a fixed set of alternatives. One can introduce such factors into the formulation in a number of ways. The simplest way is to assume that the correct response is generated by the combined hippocampal/neocortical memory with probability b_{hc}, and that on the remaining trials the animal relies on preexisting tendencies or random choice, whose probability of yielding a correct response will be denoted by b_p. The total probability of correct responding then becomes:

$$b_t(t) = b_{hc}(t) + (1 - b_{hc}(t))b_p \tag{9}$$

Although this formulation is quite minimalistic in its structure, there are several free parameters. However, the parameter R_c will be difficult to separate from the effects of the consolidation rate parameter C, and so can be set to 1 without much loss of potential expressive adequacy of the formulation. Similarly, R_h is confounded with $S_h(0)$ and can also be set to 1. In well-designed experiments there will be separate assessment of b_p or if the task is an n-alternative forced choice and the alternatives are appropriately balanced b_p will simply be $1/n$. In this case the free parameters reduce to those given in Table 1.

These equations have been fit to the data from Kim and Fanselow (1992), Zola-Morgan and Squire (1990), and the two studies that will be considered in more detail below. These fits are shown along with the data from all four studies in Figure 1. The parameters producing the fits are shown in Table 1. Fits as good as those obtained with the previously-described simulations were obtained for the Kim and Fanselow data and the Zola-Morgan and Squire data. The model also provides a moderately good fit to the data from the two other studies. We now consider these two studies in turn.

Winocur (1990). Winocur (1990) exposed rats to a conspecific demonstrator who had eaten a sample food flavored either with cinnamon or chocolate. After a delay of 0, 2, 5 or 10 days, some rats received hippocampal lesions and some received sham (control) surgery. After 10 more days for recovery, each rat was given access to both chocolate and cinnamon flavored food, and the amount of each food consumed was measured. Control rats showed a preference for the sample food eaten by the demonstrator; this preference waned over the course of about 20 days. This contrasts with the finding of Kim and Fanselow, in which there was not a significant decrement in the behavioral measure with time in their control animals. Turning to the hippocampal animals, those who were operated immediately after exposure to the demonstrator showed virtually no preference for the sample food, indicating that the initial memory trace was dependent on the hippocampus and that little or no consolidation occurred during the initial exposure event. Performance was better in the groups operated 2 and 5 days post-experience, with the 5-day hippocampals performing almost as well at test as their controls; the two ten-day groups were virtually identical and both were worse than the 5-day groups. These data suggest that in this experiment, the hippocampal trace decayed to a fairly small residual in about five days after initial exposure to the demonstrator, with a corresponding foreshortening of the duration of the consolidation interval. The data further suggest that there was considerably more decay of the neocortical traces in the Kim and Fanselow study as well.

Squire and Cohen (1979). Squire and Cohen (1979) tested retrograde amnesia in human subjects using a test based on recall for facts about television shows that aired for a single season. Their subjects were depressed humans tested either after multiple treatments of bilateral electroconvulsive therapy (ECT) or before the beginning of treatment (Control). ECT produces an amnesia in humans similar to that seen with hippocampal lesions. For present purposes, we treat ECT as equivalent to reversible removal or inactivation of the hippocampus.

Of all the human studies available, we chose to concentrate on the data from this study because the TV test presents a picture of consolidation that seems freer of contamination from intervening exposure to the material than tests based on famous faces or public events (as in Squire, Haist, & Shimamura, 1989a, and other studies); Squire and Slater (1975) went to great lengths to show that acquisition of knowledge of the TV shows depended on exposure to the shows during the year that they aired, and not on later exposure in subsequent years. It should be pointed out that the material covered by

Table 1: Parameter Values used in Fitting the Simplified Two-Memory Model to Data from Four Consolidation Experiments.

Experiment	Parameter				
	D_h	C	D_c	$S_h(0)$	$S_c(0)$
Winocur (1990)	0.250	0.400	0.075	0.900	0.100
Kim & Fanselow (1992)	0.050	0.040	0.011	0.800	0.030
Zola-Morgan & Squire (1990)	0.035	0.020	0.003	1.000	0.100
Squire & Cohen (1979)	0.001	0.001	0.001	0.500	0.000

Note: D_h represents rate of hippocampal decay; C represents rate of consolidation in off-line contexts; D_c represents rate of decay from neocortex; $S_h(0)$ represents initial strength of the hippocampal trace; $S_c(0)$ represents initial strength of the neocortical trace.

this test includes material to which the subject may have been exposed several times—if in fact the subject actually watched several episodes of each show or had secondary exposure to material about the show while it was on the air. The material tested included such things as names of characters, their roles in the show, etc. Thus this study addresses the consolidation of shared structure of the show rather than idiosyncratic details about individual scenes or episodes. Nevertheless the material is fairly idiosyncratic, in the sense that it applies to a single show and could not be derived from knowledge of other shows.

The data from the Squire and Cohen study may be misleading in one respect. Typical remote memory curves show a relatively rapid drop-off in the earlier part of the retention period with a leveling off for more remote periods (Wickelgren, 1972). The simulation, likewise, produces curves that tend to show a relatively rapid drop-off in the earlier part of the retention curve for normals, with a leveling off for more remote time periods, in accord with the typical pattern. The Squire and Cohen control data, however, show only a slight drop from the most recent to the preceding time period, with a steeper drop for later periods. Squire, Slater, and Chace (1975) report data from a study with a similar population, based on a variant of the same test, and their data actually shows slightly worse performance by depressed ECT patients for the most recent time period relative to the preceding period. Taken together the studies suggest that material from the recent past might have been less well learned relative to earlier time periods in this population of subjects. As previously noted, one possible source of this could be the subjects' severe depression during the period shortly before treatment. Depression may have affected exposure to the shows; it may have made subjects less attentive to new input than they would otherwise have been; or it may have impaired the initial formation of memory traces for the input even if attended. Such factors may be responsible for the unusual shape of the control forgetting curve—and also for some part of the apparent deficit seen in memory for the most recent time period right after ECT.

The data do, however, produce a pattern of differences between the control and ECT conditions that is comparable to those seen in the simulations. What is striking here is the fact that the pattern extends over very long periods of time relative to those obtained in the rat and monkey studies. Studies using memory for public events (Squire et al., 1989a) and for autobiographical information (MacKinnon & Squire, 1989) tend to corroborate this finding: in both cases, differences between hippocampal patients and controls are present for material more than 10 years old. This suggests hippocampal participation for over 10 years, even for material with no apparent relevance over all but the first year of this period.

Sources of Variation in Hippocampal Decay and Neocortical Learning Rate

Overall, the most striking aspect of the retrograde amnesia data presented in Figure 1 is the huge range of differences in the time-scale of the phenomenon. These differences are also reflected in the parameters of the fits to these data from the simple two-store model, as displayed in Table 1. Before we consider this issue in detail, we need to distinguish between two factors that influence the length of the consolidation interval and the outcome of the consolidation process. The first is the rate of decay from the hippocampal system, and the second is the rate of incorporation of hippocampal traces into the neocortical system. These two variables regulate the duration of the consolidation interval in contrasting ways. If the rate of decay from the hippocampus is relatively high, the consolidation period will be short because information will be lost from the hippocampus after a short period of time. If the rate of incorporation is relatively high, the consolidation period will appear short since the cortex will learn relatively quickly. In this latter circumstance we may see animals reaching ceiling levels after a relatively short consolidation interval. For the consolidation period to last a long time, both the rate of decay from the hippocampus and the rate of consolidation must be small. In general, a perusal of the parameters of the fits to the data suggests that the rate of hippocampal decay and the rate of neocortical consolidation tend to vary together in these studies. In fact, it appears that the rate of decay from the cortex also covaries with these other variables.

What might be the source of the huge differences in the time-scale of consolidation? A comparison of the Winocur (1990) and Kim and Fanselow (1992) data suggests that there are variations within species due to task differences. Some of this variation may be due to differences in the importance or salience of the information stored for the animals in these two different experiments. Winocur's animals experienced passive exposure to a conspecific who had eaten a flavored food, while Kim and Fanselow's received 15 pairings of a salient tone with painful shock. Such a salient experience may result in stronger initial hippocampal traces that show greater resistance to decay. In this connection it is interesting to note that the decay of plastic changes in the hippocampus does vary as a function of the magnitude of the inducing stimulation. Barnes (1979) has found that LTP produced by relatively weak inducing stimuli decays over a period of days while LTP produced by more massive or repeated stimulation can last for weeks, and Abraham and Otani (1991) review a number of subsequent studies producing decay rates falling into two groups, one, produced with fewer or weaker inducing stimuli, showing a decay rate of about 0.28 per day and another other, produced with more or larger inducing stimuli, showing a rate of about 0.05 per day. These rates closely parallel the hippocampal decay rates shown in Table 1 for our fits to Winocur (1990) and Kim and Fanselow (1992) respectively. Interestingly, Sutherland (personal communication) has repeated the Kim and Fanselow (1992) study varying the number and intensity of foot shocks and finds that the persistence of the memory for spatial context is reduced with fewer pairings. In natural situations differences in extent of initial learning and/or magnitude of decay might be induced by neuromodulatory systems activated in strongly motivated or emotionally charged situations (Gold & McGaugh, 1984).

Our view of the purpose of neocortical learning—to foster the gradual discovery of shared structure—motivates two other sorts of suggestions about the possible sources of the differences observed between the different experiments. One possibility is that there may be species differences in the rate of neocortical learning, arising from different evolutionary pressures. In animals with relatively short life-spans, a very slow rate of neocortical learning would make little sense, since the animal's life could be over before much adaptation has taken place. Furthermore, if the structure such animals needed to extract from the environment through experience was relatively straightforward, it might be possible for the neocortical systems of these animals to learn it relatively quickly. On the other hand, in animals with much longer life spans—especially in humans who must master complex bodies of knowledge that vary from culture to culture—it may be that incorporation of new knowledge into the neocortical system must occur at a much slower rate. The extreme slowness of learning even often-repeated aspects of post-lesion experience in profound human amnesics (Milner et al., 1968) may be due, at least in part, to the necessity of extremely small learning rates in humans. These considerations lead to the possibly counter-intuitive prediction that hippocampally damaged rats might show faster acquisition than comparably damaged primates on tasks dependent on learning in the neocortical system.

A second possibility is that there are age differences in the rate of neocortical learning. Changes in the rate of neocortical learning with age could be one of the reasons why consolidation appears slower in the human studies than in the rat or monkey research. All of the systematic human data that we know of comes from adults; the monkeys and rats used in other studies would have been considerably younger in raw chronological age. There is relatively little evidence directly related to this age hypothesis, though Squire (1992) makes one suggestive observation: He notes that the retrograde amnesia in patient HM may have been somewhat shorter—only about three years— compared to the retrograde amnesia of about 10 years seen in the older group of patients tested by MacKinnon and Squire (1989). This sort of difference would be expected if the rate of consolidation were to decrease gradually with increasing age.

Why might the rate of neocortical learning change with age? One functional reason for this arises from a consideration of the optimal procedure for estimating population statistics in online statistical estimation procedures. In general, in these procedures, it is best to make relatively large adjustments in response to initial observations, and then gradually reduce the size of the adjustments as the sample size increases (White, 1989). For example, the optimal on-line procedure for estimating a simple statistic, such as the mean of a number of observations, is to adjust the estimate after each observation by an amount equal to one over the total number of observations taken, including the last:

$$\Delta e_n = \frac{1}{n}(o_n - e_{n-1}) \qquad (10)$$

In this case e_n, the estimate of the population mean after the nth observation o_n, is always exactly equal to the mean of the n sample observations. This procedure yields the optimal, unbiased estimate of the population mean based on the entire preceding sample at every step. In more complex networks it is not necessarily the case the one should begin immediately to reduce the learning rate, but convergence to the population value of a statistic that is being estimated through a stochastic learning procedure generally requires the use of a gradually diminishing learning rate (Darken & Moody, 1991).

Given this observation, it may make sense to begin life using relatively large neocortical learning rates, to begin extracting structure from experience relatively quickly, then to reduce the learning rate gradually as experience accumulates. This statistical argument may be a part of the functional explanation for various critical period phenomena in development. If this is correct, we would expect to see much more rapid acquisition of the shared structure of events and experiences in younger human amnesics and animals with hippocampal lesions, relative to older amnesic groups.

One consideration relevant to the points raised in this section is the fact that that the effective learning rate of a system can vary as a function of factors other than just the learning-rate parameter of the synaptic modification rule. In particular, networks that have developed strong connection weights after a period of learning can exhibit a lack of sensitivity to new information (Munro, 1986; Miller et al., 1989). This suggests that reduction in the effective learning rate with age could be a byproduct of previous learning.

The preceding discussion relates to the consolidation rate parameter in our simple model (actually the product of the neocortical learning rate and the reinstatement rate), but should not be construed as providing a basis for predicting longer hippocampal persistence of memories with age. In fact, it may be that initial storage of information in the hippocampus gets poorer with age and/or that the decay rate increases. Barnes and McNaughton (1985) have found evidence that older rats exhibit more rapid decay of hippocampal LTP and also exhibit faster forgetting of spatial information. Perhaps the effective rate of decay of information from the hippocampal system increases with age even as the rate of neocortical learning decreases. If so, the separate changes would compound each other's effects, thereby doubly diminishing the plasticity of the aging neocortical system. Obviously, the matters raised here deserve considerably more exploration. It would be useful to know much more about how consolidation changes as a function of species, age, task variables, and prior learning.

Infantile amnesia. The fact that the neocortical learning rate may be relatively high early in life, before settling down to relatively lower rates as more and more structure is extracted, may provide at least a partial account of the phenomenon of infantile amnesia: the fact that humans have little or no explicit memory from the earliest periods of their lives. A recent review by Howe and Courage (1993) concludes that infantile amnesia cannot be explained away as a simple result of immaturity of the nervous system. We suggest that the phenomenon may be due instead to rapid initial change in the structured representations used in the neocortical system. Hippocampal traces based on immature representational systems would be difficult to access, since the cortical representation of an appropriate probe would have changed. They would also be more difficult to interpret if reinstated, since the reinstated representations would no longer make sense in terms of the more mature system of neocortical representations.

Fast learning outside the hippocampal system. In animal studies, even complete lesions to the hippocampal system can leave certain forms of learning intact. As Rudy and Sutherland (1994) point out, spared learning in such cases appears to depend on the formation of associations of a simple, discrete cue with an unconditional stimulus or the availability of reinforcement. Simple correlations between discrete cues and outcomes are a kind of low-order structure that is shared across situations, and therefore it makes sense for such correlations to be learned in the neocortical system. Much of the data could be explained if, as suggested above, the neocortical learning rate is relatively high in rats, and if performance depends on the use of these neocortical associations rather than the conjunctive representations formed in the hippocampus. Hippocampal conjunctive representations might be of little use in tasks that require the generalization of a response to a specific cue in a new context, rather than the repetition of a response acquired in a specific old context (Rudy & Sutherland, 1994).

It may seem to strain our basic approach, however, that learning of simple associations can sometimes occur in just one or a few trials; according to our analysis, acquisition of such correlations by the neocortical system should occur somewhat gradually, even in animals like rodents, to prevent potential interference with previously-established associations to the same stimuli. From a functional point of view, however, it seems reasonable to suppose that evolution might provide mechanisms that over-ride this consideration in situations where very rapid adaptation may be necessary for survival. It is well-accepted that there are special systems for rapid learning of certain types of associations, e.g., between the taste of food and sickness (Garcia, Ervin, & Koelling, 1966), that appear to be specific to the survival needs of particular species, and it obviously makes sense for such systems to be able to learn quickly. Such systems might be seen as providing a second form of fast learning, quite different from the hippocampal system in many respects, but similar to it in providing a mechanism for rapid learning that leaves the neocortical system free for the gradual discovery of shared structure.

There is another basis within our framework for understanding relatively spared learning of simple cue-outcome contingencies even in cases where these are not strongly survival related. The idea is closely related to our earlier observation that one can optimize both initial and final performance by making large initial weight changes and then gradually reducing the sizes of these changes. A variation on this would be to learn simple, first-order relationships relatively quickly (especially if the stimuli can be tagged as novel so that prior associations would not be a consideration), while extracting higher-order relationships at a lower rate. This makes sense from a statistical perspective since simple first-order relations are generally apparent in much smaller samples of data than higher-order relationships. If the neocortex exploited this strategy, we would expect learning of simple associations to be relatively spared following hippocampal lesions compared to learning of more complex relationships. Indeed, as noted earlier, hippocampal lesions do produce selective deficits in negative patterning and other paradigms that require the animal to master higher-order cue-outcome contingencies. In this context, it is interesting that hippocampal system lesions only produce reliable impairments of learning of higher-order relationships in paradigms where these relationships are pitted against simple first-order relationships (Rudy & Sutherland, 1994). Negative patterning is such a case, since the animal must learn not to respond to the conjunction of A and B, even though responses to A and B are each individually reinforced. On the other hand, when only compound stimuli are used, and the animal must respond to the compounds AC and BD but not to the compounds AD and BC, there are no competing first-order relationships, and hippocampal rats appear relatively unimpaired in this case. The data are consistent with the idea that structures outside the hippocampal system in the rat can learn higher-order associations, but that the strength of these associations builds more gradually than the strength of first-order associations.

General Discussion

We have presented an account of the complementary roles of the hippocampal and neocortical systems in learning and memory, and we have studied the properties of computational models of learning and memory that provide a basis for understanding why the memory system may be organized in this way. We have illustrated through simple simulations how we see performance and consolidation arising from the joint contributions of the hippocampal system and the neocortical system, and we have considered why there may be variation in learning rate as a function of age, species, and other functional considerations. In this section, we compare the approach we have taken to some other views of the role of the hippocampal system in learning and memory. It is beyond the scope of this paper to offer an exhaustive summary and comparison of the present theory to other views, but there are a few major points of similarity and difference that warrant discussion.

Perspectives on Retrograde Amnesia

Our treatment of the complementary roles of the hippocampal and neocortical systems rests on the centrality of the phenomenon of temporally graded retrograde amnesia. The phenomenon calls out for a theory that specifically accords the hippocampal system a relatively extended, but nevertheless time-limited role, in some but not all memory tasks. In this respect our treatment continues a theme that was emphasized in some of the earliest discussions of amnesia (e.g., Ribot, 1882). The notion that the hippocampal system plays a role in consolidation began to emerge with the initial studies of HM (Scoville & Milner, 1957; Milner, 1966). It was adopted in the theoretical proposals of Marr (1971) and has been strongly emphasized in the work of Squire and his collaborators over a twenty-year period (Squire et al., 1975; Squire et al., 1984; Alvarez & Squire, 1994).

Squire et al. (1984) treat temporally graded retrograde amnesia as a reflection of a gradual process of memory reorganization. Our proposals accord with, and elaborate, this suggestion. The models we have presented produce such reorganizations, and our analysis of these models provides an explicit account of the reasons why these reorganizations should necessarily be slow that is not present in the Squire et al. (1984) account. Our proposals also build on the earlier ideas of Marr (1970, 1971). He saw consolidation as a process of sorting experiences into categories, and noted that this sorting process would require an adequate statistical sample of the environment. This proposal is a specific example of our more general claim that the neocortical system is optimized for the discovery of the shared structure of events and experiences.

The idea of consolidation as the reflection of a process in which the hippocampus plays back information to the neocortex may have originated with Marr (1971) as well. He proposed that the hippocampal system stored experiences as they happened during the day, and then replayed the memories stored in the hippocampal system back to the neocortex overnight, to provide data for the category formation process as he envisioned it. Several neurophysiologists (McNaughton, 1983; Pavlides & Winson, 1989; Buzsaki, 1989; Wilson & McNaughton, 1993) have pursued this idea. McNaughton (1983) suggested that the high-frequency, complex-spike burst discharges that occur during hippocampal sharp waves are one source of such reinstatement. This idea has been elaborated in considerable detail by Buzsaki (1989). The idea of the hippocampus as providing extra learning opportunities for the neocortex has also been proposed by Milner (1989). We first discussed the idea in McClelland, McNaughton, O'Reilly, and Nadel (1992), and it has recently been adopted by several other authors (Alvarez & Squire, 1994; Lynn, 1994; Treves & Rolls, 1994). Earlier versions of our simulations of consolidation in the studies of Kim and Fanselow (1992) and Zola-Morgan and Squire (1990) were presented in McClelland, McNaughton, and O'Reilly (1993). In modeling studies contemporaneous with and independent of ours, Alvarez and Squire (1994) developed a simple neural network model of the hippocampal and neocortical systems and showed how it could capture the general form of the consolidation functions shown in Figure 1, and Lynn (1994) developed a simple conceptual model quite similar to the one we have presented.

Many theorists have focused primarily on the anterograde effects of hippocampal lesions. Many of these theoretical discussions have suggested that the hippocampal system directs the choice of representations in neocortex; the neocortical system is treated as an impoverished learning device that needs the hippocampus to function more effectively in certain contexts. This is a rather different form of the idea of the hippocampus as teacher to the neocortex than the one that we have proposed, since our idea is simply that the hippocampus provides training trials, allowing the cortical system to select representations for itself through interleaved learning. Several variants of the idea that the hippocampus directs or influences neocortical representations have been proposed. (Wickelgren, 1979; Rolls, 1990) have suggested that the hippocampus is necessary to assign distinct cortical representations to particular novel conjunctions of inputs, so that the neocortex can treat these separately from other overlapping episodes and events. Rudy and Sutherland (1994) suggest that the hippocampus may increase the salience of neocortical representations of cue conjunctions, facilitating the learning of conjunctive relationships, and Schmajuk and DiCarlo (1992) and Gluck and Myers (1993) assume that the hippocampus provides error signals that direct the neocortex in the formation of representations of cue combinations. In most of these models, the hippocampus plays its role at the time of initial memory formation, leaving no basis for expecting any retrograde amnesia. However, Wickelgren (1979) suggested that the hippocampus was necessary for the initial selection of the neocortical representation and for its subsequent reactivation, until direct intracortical connections can become established (through gradual learning). Treves and Rolls (1994) provide an extension of the theory of Rolls (1990) that encompasses essentially the same idea. The result is that these theories can provide an account for the phenomenon of temporally graded retrograde amnesia. The main difference is that our approach provides a principled, functional reason why the neocortex should necessarily learn gradually—and thus that retrograde amnesia should necessarily be temporally extended—-while these other approaches do not.

Several other authors have proposed that the hippocampus is necessary for a particular type of information processing or representation that is crucial for some memory tasks. For example, several authors distinguish between pathway-based learning, in which modifications occur directly in pathways involved in specific acts of information processing, and more cognitive forms of learning associated with performance in explicit memory tasks. This or a related distinction may be found in several other places (Squire, 1992; Humphreys, Bain, & Pike, 1989; O'Keefe & Nadel, 1978; Mishkin, Malamut, & Bachevalier, 1984; Cohen & Eichenbaum, 1993; Warrington & Weiskrantz, 1978). A related distinction is made in our approach as well, though we differ from some of these other theorists in one crucial respect: We emphasize the fact that ultimately, both forms of learning can occur in the neocortical system. Once again, it is the phenomenon of temporally graded retrograde amnesia that is crucial for our theory. Those who view the hippocampus as necessary for a specific type of representation, storage, or information processing that is viewed as crucial for performance in certain memory tasks appear to predict that retrograde amnesia will affect material from all past time periods, and will not be time-limited.

In summary, three different kinds of roles have been suggested for the hippocampus: One kind has it aiding the cortex in selecting a representation to use at the time of storage. Another type has it providing a crucial form of representation (or learning, or processing) not available to the neocortex, that is necessary for performance in certain sorts of memory tasks. The third type of theory has the hippocampus playing an explicitly time-limited role in the formation of neocortical representations. The first type of theory can explain anterograde amnesia, but appears to offer only *ad-hoc* accounts of retrograde amnesia. The second type of theory can explain retrograde amnesia as well as anterograde amnesia, but appears to predict that retrograde amnesia will not be temporally graded. While aspects of all three types have considerable appeal, only the third type of theory offers a principled account of temporally graded retrograde amnesia. Of theories of the third type, ours is the first to offer an explicit computational account of why the period of hippocampal involvement must necessarily be temporally extended.

Other Points of Comparison

Several additional features of other approaches to the roles of the neocortical and hippocampal systems in learning and memory warrant consideration in comparison with our proposals. At first glance our approach may seem to contrast with some of these other approaches but on closer inspection many of these approaches may be complementary, and many apparent differences may be matters of emphasis and perspective.

Hippocampal conjunctive coding, spatial representation, and the temporally extended role of the hippocampal system during consolidation. A number of investigators have proposed that the hippocampus plays a special role in learning contingencies involving conjunctions of cues (Wickelgren, 1979; Sutherland & Rudy, 1989; Rolls, 1990), and both Schmajuk and DiCarlo (1992) and Gluck and Myers (1993) have proposed explicit computational models in which the hippocampal system plays a crucial role in the formation of internal representations of cue combinations. Our account of the role of the hippocampus in learning and memory is similar in that it assumes that the hippocampal system is necessary for the rapid formation of conjunctive representations, but differs from these other proposals in that we assume these representations are initially formed in the hippocampal system. The fact that the Schmajuk and DiCarlo (1992) and Gluck and Myers (1993) models both account for a range of phenomena from the complex literature on classical conditioning suggests that the essential computational properties of these models may have considerable validity.

Rudy and Sutherland (1994), who have stressed the role of the hippocampal system in memory that depends on cue conjunctions, have suggested that this may be the basis, at least in part, for its special role in spatial navigation, place learning, and conditioning involving the learning context as a discriminative cue. By the same token, O'Keefe and Nadel (1978) proposed that mechanisms initially derived for spatial representations and processes may be recruited for other functions and McNaughton et al. (1989) have argued that spatial learning may involve specific conjunctions of locations and movements. Whether the hippocampal system is primarily or was initially specifically a spatial processing system, as O'Keefe and Nadel (1978) have argued, or is essentially a system designed specifically to handle cue conjunctions, as proposed by Rudy and Sutherland (1994), may be undecidable and even largely irrelevant to considerations about function, given that neural structures often serve multiple functions and are often recruited to new functions through evolution (Rozin, 1976; Gould & Lewontin, 1979).

It seems to us that the conjunctive coding perspective, the spatial learning perspective, and the perspective taken here are not mutually exclusive: It may be best to view all of these functions as synergistic. We offer two illustrations of this point. First, consider the possible synergy between a spatial memory system and context sensitivity. Because space provides a contextual framework for all experiences, it would not be surprising if the hippocampal system evolved certain intrinsically spatial mechanisms to facilitate the linking of experiences to their spatial contexts. At least part of the evidence from recordings of hippocampal neurons *in vivo* is compatible with the existence within its synaptic matrix of intrinsic associative links between adjacent points in space (Knierim et al., in press). These links would provide a preconfigured substrate to which specific objects and events might become bound Gothard et al. (1994). This would facilitate the construction of a composite memory incorporating events occurring at different times in the same spatial location (Rawlins, 1985). Second, consider the possible synergy between the use of sparse, conjunctive representations in the hippocampus and the temporally extended role it has to play according to our theory of consolidation. Hippocampal representations must be maintained for an extended period of time to ensure adequate consolidation, even as new memories are continually being added. Since the knowledge that maintains these representations is stored in connections among units active in each representation, it is crucial to minimize the overlap of the representations to minimize interference; otherwise the hippocampal memory system would itself suffer from catastrophic interference and few representations would remain for the long term. Thus, it is possible to see the special role of the hippocampus in learning that depends on cue conjunctions—including memory for the contents of specific episodic experiences and spatial or contextual learning—as a reflection of a coding style that serves at least as much to minimize interference among sequentially stored memories as to provide a basis for conjunctive (or equivalently spatial, episodic) learning *per se*.

Explicit and declarative memory. Schacter (1987, 1994) has stressed the descriptive value of the distinction between *explicit* and *implicit* memory in characterizing aspects of the human learning and memory literature. He defines explicit memory tasks as tasks that require the deliberate or conscious access to prior experience, whereas implicit memory tasks are those that do not require such deliberate or conscious access to prior experience. Human amnesics are clearly impaired in explicit memory tasks on this definition. However, amnesics are also impaired in acquisition of new, arbitrary factual information, whether or not the use of this information is accompanied by deliberate or conscious recollection of previous experiences in which this information was presented (Shimamura & Squire, 1987). (The evidence was considered in paragraph (2) under *The Role of the Hippocampal System in Learning and Memory*). Thus, it does not appear that amnesia is a deficit that is restricted to explicit memory as defined by Schacter (1987). Squire (1992) maintains the view that lesions to the hippocampal system produce a deficit in "declarative memory" by which he means memory whose *contents* can be consciously brought to mind or declared. This term does encompass fact memories of the form studied by Shimamura and Squire (1987), but uncertainties remain concerning whether human amnesics' deficits are restricted to tasks involving memories that are consciously accessible. While amnesics are impaired in stem-completion tests of sensitivity to novel associations Schacter and Graf's (1986), it is unclear to what extent conscious accessibility is necessary for sensitivity to new arbitrary associations in such paradigms.

Our perspective gives us a different vantage point on this issue. We have adopted the view that it is the rapid formation of novel, conjunctive associations that crucially depends on an intact hippocampal system, and we would suggest that the forms of memory that are encompassed by the terms "explicit" and "declarative" are good examples of forms of memory that depend on the rapid formation of such associations, but are not necessarily the only ones. Other forms of memory that are not explicit or declarative in any obvious sense might depend on the rapid formation of such associations as well.

Flexible use of memory. The concepts of explicit and declarative memory are even more difficult to operationalize for animal studies than they are for humans. However, Cohen and Eichenbaum (1993) have suggested that the hippocampus may be specialized for the representation of recent memories in a form that supports their flexible use (see also Eichenbaum et al., 1994). This is an attractive idea in that the human ability to report declaratively on the contents of a recent experience might be treated as just one example of flexible use, and Cohen and Eichenbaum's proposal is reminiscent of the distinction between memories and habits introduced by Mishkin et al. (1984). In our view, the flexible use of recent memory is not a unitary function of the hippocampal system but depends on cooperation of the hippocampus and other brain systems, particularly the frontal lobes. Cohen and O'Reilly (in press) present this idea, and suggest that the role of the hippocampal system is to provide for the rapid auto-associative storage of the arbitrary contents of particular episodes and events, allowing for their reconstruction via the associative pattern reinstatement process we have repeatedly discussed. The other parts of the system (e.g., the prefrontal cortex) can influence hippocampal recall by providing different activity patterns as cues to the hippocampus, and are in turn influenced by the information that is thereby recalled. These interactions would then lead to the flexible use of information stored in the hippocampus.

Reference versus working memory. The proposal of Olton, Becker, and Handelmann (1979) that the hippocampal system is necessary for what they called working memory (memory for recent information of specific current relevance) but not reference memory (memory for invariant aspects a task situation) bears some similarity to our view that the cortex is specialized for the gradual discovery of the shared structure of events and experiences, while the hippocampus is necessary for the rapid storage of the contents of specific episodes and events. Where we differ from Olton et al. (1979), however, is in suggesting that the hippocampus can support relatively rapid acquisition of all aspects of a particular experience. Those aspects that are invariant in a task situation could guide initial task performance, prior to neocortical consolidation. On this basis we would expect that hippocampal system lesions would affect the initial acquisition of spatial reference memory, but not performance based on learning occurring prior to the lesion. In fact these expectations are borne out in the literature. Barnes (1988) reviews several spatial working memory studies and concludes that when the lesion occurs before training there is invariably a marked impairment. However when the lesion occurs after training there may be little or no impairment, and Barnes (1988) suggests that the variability may be due at least in part to differences in the delay between initial learning and testing. In support of this she cites a study by Sutherland, Arnold, and Rodriguez (1987) in which animals with dentate gyrus lesions showed dramatic impairments on a spatial reference memory task when the lesion occurred shortly after initial acquisition, but not when the lesion occurred after a delay of several weeks.

Binding. It has often been suggested that the hippocampal system provides a mechanism that binds together the diverse aspects of the cortical representation of a specific episode or event. Variants of this idea can be found in Wickelgren (1979), Squire et al. (1984), Teyler and Discenna (1986) and Damasio (1989). Some of these proposals—most explicitly, the one by Teyler and Discenna (1986)—suggest that the hippocampal system does not store the memory itself, but rather stores only a list of addresses of or pointers to the diverse locations in the neocortex where the memory itself is stored. Since we suggest that the plastic changes responsible for the initial storage of the contents of particular episodes and events take place within the hippocampal system, our view may seem at first glance to contrast sharply with the view of the hippocampal representation as a list of addresses bound together. However, closer scrutiny reveals that our view may be more similar to the Teyler and Discenna (1986) view than is initially apparent (see Alvarez & Squire, 1994, for a related argument). As noted previously, our proposal does not require that somehow a full copy of the neocortical pattern of activation is transferred to the hippocampal system; rather we assume the hippocampus uses a compressed representation, and simply needs to encode enough information about the pattern for the neocortex to reconstruct enough of it to guide overt responses. This idea of the hippocampal system working with compressed representations can be seen as similar to the Teyler and Discenna (1986) proposal, replacing their *addresses* with our *compressed representations*. A further point of similarity arises from the fact that additional knowledge is needed to implement the pattern compression and decompression processes. This knowledge is to be found in the connections within the cortical system and in the connections leading to and from the hippocampal system from the neocortical system. Compression is carried out by the connections leading into the hippocampal system, resulting in a reduced representation in the entorhinal cortex, the gateway to the hippocampus itself. This reduced representation is then the one that is stored in the hippocampal system. When this representation is retrieved at a later time, return connections from the entorhinal cortex to the neocortical system, as well as connections within the neocortex, participate in the reinstatement of the neocortical pattern that was present at the time of storage. This proposal shares with the proposals of Teyler and Discenna (1986) and others the idea that much of the information needed to reconstruct a particular pattern of activation is not stored in the hippocampal system.

Teyler and Discenna's proposal includes the suggestion that the actual content of hippocampal-system dependent memories is stored within local circuits in the neocortex at the time of learning. Squire et al. (1984) raise this possibility as well. The idea appears to be that patterns of activation in local circuits are stored via plastic changes that occur within these circuits during the initial experience, and that the hippocampus only binds these local patterns together so that the local pattern in one part can be reactivated by patterns arising in other parts. The plastic changes in these local circuits constitute, on this view, the extra information needed to turn the addresses stored in the hippocampus into a neocortical memory trace.

This aspect of Teyler and Discenna's (1986) proposals does contrast with our proposals, since we have suggested a dissociation between the hippocampal system and the neocortical system based on fast versus slow learning, whereas Teyler and

Discenna (1986) appear to be suggesting that there must be some fast learning within the neocortex. However, it would be possible to reconcile our proposals with theirs without abandoning our fundamental claim that there must be a division of labor between fast and slow learning systems, by revising the placement of the anatomical boundary between the fast and slow learning systems. One specific possibility is that the superficial layers of the neocortex are part of the fast learning system, and that the slow-learning system is located in deeper layers. On this view the fast learning system is anatomically quite distributed, with the hippocampus itself serving as the system's "convergence zone" Damasio (1989). There are contrasts between the superficial layers of the neocortex (layers two and three) and deeper layers that are consistent with this suggestion. There is a higher density of NMDA receptors in superficial layers of neocortex (Monaghan & Cotman, 1989), and recent studies of two areas of neocortex indicate that the superficial layers use sparser representations (Skaggs, McNaughton, Barnes, Moore, Duffield, Frank, & D'Monte, 1994). Also, it is the superficial layers of neocortex that exchange bi-directional connections with the input-output zones of the hippocampal system. This suggestion brings our theory into much closer alignment with the suggestions of Teyler and Discenna (1986) and others who attribute a binding function to the hippocampus. The suggestion also aligns our approach better with proposals that the hippocampus may contribute to the assignment of neocortical representations at the time of initial learning, if we stipulate that the neocortical representations in question are specifically the ones used in the superficial layers. Given the convergence of inputs to the hippocampus and the divergence of the return projections, the presence of the hippocampus could well influence the representations used in the superficial layers (Rolls, 1990; Treves & Rolls, 1994).

There is, however, and alternative to accepting the idea that substantial changes must be made in the neocortex at the time of initial memory formation to allow for decoding of the compressed hippocampal representation. The alternative is the idea that the knowledge needed for the encoding and decoding operations is already present in the connections into and out of the hippocampal system and in the intracortical connections that participate in the reinstatement process. If these connections were part of the slow-learning neocortical system, their weights would come to exploit the structure (redundancies and constraints) characteristic of the ensembles of events previously experienced, enabling successful compression and decompression of new patterns that exhibit the same structure, but not of unstructured patterns or patterns exhibiting unfamiliar structure. This is, in fact, exactly what happens in the connectionist pattern compression models mentioned earlier in this article. An appealing feature of this proposal is that it would contribute to stability of the bi-directional mapping of patterns of activation between the fast and slow learning systems, since the bi-directional connections between the hippocampus and neocortex would not themselves be subject to rapid changes during the storage of new associations. Rapid change in these bi-directional pathways would tend to interfere with the reinstatement of older hippocampal memories. Since in the human case it appears that the hippocampal system can contribute to reinstatement for a period of years after initial memory acquisition, it would be desirable to avoid such interference. The proposal that the bi-directional connections between the hippocampal and neocortical systems exploit the structure present in ensembles of experiences predicts that hippocampal memory will be far superior for materials conforming to the structural constraints that were embodied in the entire ensemble of past experiences, than it will be for totally random patterns or patterns that embody quite different constraints, since in the latter cases the knowledge built into the connection weights would not be applicable, and encoding and/or decoding would fail. Indeed, one of the most pervasive findings in the human memory literature is the finding that memory is far better when the material to be learned conforms to familiar structures (Bartlett, 1932).

Both of the ideas considered in the last two paragraphs have considerable appeal. However, both raise a host of further issues as well, and at the present we see no clear basis for choosing between them. Stepping back from these specific matters, though, our comparison of our views with those of Teyler and Discenna (1986) underscores an important general point: Apparent differences between theories may not reflect a fundamental incompatibility at a functional level. We have already seen in previous sections that some apparent contrasts may reflect differences of focus and emphasis, and the present section (as well as the next section) further exemplifies this point. In addition, we now can see that other apparent contrasts might hinge not so much on differences in claims about the functional organization but on different assumptions about the alignment of the functional organization with the neural substrate or on differences in specific aspects of the proposed implementation.

Prediction. The last perspective we will consider on the role of the hippocampus is the idea that it is necessary to predict the future based on the present and the recent past. This kind of suggestion has been made by several authors (Levy, 1989; Schmajuk & DiCarlo, 1992; Gluck & Myers, 1993). We agree that prediction based on recent experience is impaired after damage to the hippocampal system. However, we view prediction as among the many special cases of the associative learning that occurs in the hippocampus. Prediction can arise from associative storage and subsequent retrieval through pattern completion. One possibility is that the pattern of activation produced in the hippocampal system at any point in time reflects experience over a small temporal window. Auto-associative storage of this pattern in the hippocampal system would then link the situation, action, and outcome together into a single memory trace. At a later time, when the beginning of a previously-experienced sequence occurs, this could serve as a probe to the hippocampal system, and pattern completion would then allow reinstatement of the next step or steps in the sequence. This idea that the pattern of activation at a particular point in time actually encompasses some temporal window could be coupled with the assumption that the pattern is not associated with itself, but with a pattern arising at a slightly later time (Levy, 1989; Minai & Levy, 1993; Larson & Lynch, 1986; McNaughton & Morris, 1987; Gluck & Myers, 1993). This hybrid scheme would permit recall of temporal sequences as well as auto-associative completion of material in the overlapping patterns that are active at adjacent times.

Conclusion

In this article, we have treated the phenomenon of consolidation as a reflection of the gradual incorporation of new knowledge into representational systems located primarily in the neocortical regions of the brain. Our proposal has its roots in the work of Marr (1970, 1971) and Squire et al. (1984), but we have given it a clearer computational motivation than these earlier investigators, and we have pointed to computational mechanisms that indicate how the incorporation of new knowledge can gradually cause the structure itself to adapt. Nevertheless our analysis is far from complete, and many details of the implementation and physiological realization of the complementary learning systems we have proposed remain obscure. Our analysis may address the two key questions we have posed in this article, but in so doing it raises many new ones. Answering these questions will depend on the emerging synthesis of computational, behavioral, and neurophysiological investigation.

References

Abraham, W. C., & Otani, S. (1991). Macromolecules and the maintenance of long-term potentiation. In F. Morrell (Ed.), *Kindling and synaptic plasticity* (pp. 92–109). Boston: Birkhauser.

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science*, *9*, 147–169.

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

Barnes, C. A. (1979). Memory deficits associated with senescence: A neurophysiological and behavioral study in the rat. *Journal of Comparative and Physiological Psychology*, *93*, 74–104.

Barnes, C. A. (1988). Spatial learning and memory processes: The search for their neurobiological mechanisms in the rat. *Trends in Neurosciences*, *11*(4), 163–169.

Barnes, C. A., Jung, M. W., McNaughton, B. L., Korol, D. L., Andreasson, K., & Worley, P. F. (1994). Ltp saturation and spatial learning disruption: Effects of task variables and saturation levels. *Journal of Neuroscience*, *14*, 5793–5806.

Barnes, C. A., & McNaughton, B. L. (1985). An age comparison of the rates of acquisition and forgetting of spatial information in relation to long-term enhancement of hippocampal synapses. *Behavioral Neuroscience*, *99*, 1040–1048.

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

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

Barrionuevo, G., & Brown, T. H. (1983). Associtive long-term synaptic potentiation in hippocampal slices. *Proceedings of the National Academy of Sciences*, *80*, 7347–7351.

Bartlett, F. C. (1932). *Remembering*. Cambridge, MA: Cambridge University Press.

Barto, A. G., & Jordan, M. I. (1987). Gradient following without backpropagation in layered networks. *Proceedings of the IEEE International Conference on Neural Networks*, Vol. 2 (pp. 629–636). San Diego: SOS Printing.

Barto, A. G., Sutton, R. S., & Brouwer, P. S. (1981). Associative search network: A reinforcement learning associative memory. *Biological Cybernetics*, *40*, 201–211.

Bliss, T. V. P., & Gardner-Medwin, A. R. (1973). Long-lasting potentiation of synaptic transmission in the dentate area of the unanaesthetized rabbit following stimuliation of the perforant path. *Journal of Physiology (London)*, *232*, 357–371.

Bliss, T. V. P., & Lømo, T. (1973). Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimuliation of the perforant path. *Journal of Physiology (London)*, *232*, 331–356.

Bowers, J. S., & Schacter, D. L. (1993). Implicit memory and test awareness. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *16*, 404–416.

Buzsaki, G. (1989). Two-stage model of memory trace formation: A role for 'noisy' brain states. *Neuroscience*, *31*, 551–570.

Chrobak, J., & Buzsaki, G. (1994). Selective activation of deep layer (V-VI) retrohippocampal neurons during hippocampal sharp waves in the behaving rat. *Journal of Neuroscience*, *14*, 6160–6171.

Cleeremans, A. (1993). *Mechanisms of implicit learning: Connectionist models of sequence processing*. Cambridge, MA: MIT Press.

Cohen, J. D., & O'Reilly, R. C. (in press). 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*. Hove, England: Earlbaum.

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

Cohen, N. J., & Squire, L. R. (1980). Preserved learning and retention of pattern analyzing skill in amnesia: Dissociation of knowing how and knowing that. *Science*, *210*, 207–209.

Collingridge, G., Kehl, S., & McLennan, H. (1983). Excitatory amino acids in synaptic transmission in the Schaffer collateral-commisural pathway of the rat hippocampus. *Journal of Physiology (London)*, *334*, 3–46.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*, 240–247.

Cottrell, G. W., Munro, P., & Zipser, D. (1987). Learning internal representations from gray-scale images: An example of extensional programming. *Proceedings of the 9th Annual Conference of the Cognitive Science Society* (pp. 462–473). Hillsdale, NJ: Erlbaum.

Damasio, A. R. (1989). The brain binds entities and events by multiregional activation from convergence zones. *Neural Computation*, *1*, 123–132.

Darken, C., & Moody, J. (1991). Note on learning rate schedules for stochastic optimization. In R. P. Lippman, J. E. Moody, & D. S. Touretzky (Eds.), *Advances in Neural Information Processing Systems 3* (pp. 832–838). Palo Alto: Morgan Kaufmann.

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*, *107*, 227–234.

Douglas, R. M. (1977). Long-lasting potentiation in the rat dentate gyrus following brief, high-frequency stimulation. *Brain Research*, *126*, 361–365.

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

Fahlman, S. E. (1981). Representing implicit knowledge. In G. E. Hinton, & J. A. Anderson (Eds.), *Parallel models of associative memory* (Chap. 5, pp. 145–159). Hillsdale, NJ: Erlbaum.

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

French, R. M. (1991). Using semi-distributed representations to overcome catastrophic forgetting in connectionist networks. *Proceedings of the 13th Annual Cognitive Science Conference* (pp. 173–178). Hillsdale, NJ: Erlbaum.

French, R. M. (1992). Semi-distributed representations and catastrophic forgetting in connectionist networks. *Connection Science*, *4*(3-4), 365–377.

Gaffan, D., & Murray, E. A. (1992). Monkeys (macaca fascicularis) with rhinal cortex ablations succeed in object discrimination learning despite 24-hour intertrial intervals and fail at matching to sample despite double sample presentations. *Behavioral Neuroscience*, *106*, 30–38.

Galland, C. C., & Hinton, G. E. (1991). Deterministic Boltzmann learning in networks with asymmetric connectivity. In D. S. Touretzky, J. L. Elman, T. J. Sejnowski, & G. E. Hinton (Eds.), *Connectionist Models: Proceedings of the 1990 Summer School* (pp. 3–9). San Mateo, CA: Morgan Kaufmann.

Garcia, J., Ervin, F. R., & Koelling, R. A. (1966). Learning with prolonged delay of reinforcement. *Psychonomic Science*, *5*, 121–122.

Gilbert, C. D. (1994). Dynamic properties of adult visual cortex. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 73–90). Cambridge, MA: MIT Press.

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

Glisky, E. L., Schacter, D. L., & Tulving, E. (1986b). Learning and retention of computer-related vocabulary in memory-impaired patients: Method of vanishing cues. *Journal of Clinical and Experimental Neuropsychology*, *8*, 292–312.

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

Gold, P. E., & McGaugh, J. L. (1984). Endogenous processes in memory consolidation. In H. Weingartner, & E. S. Parker (Eds.), *Memory consolidation: Psychobiology of cognition* (Chap. 3, pp. 65–83). Hillsdale, NJ: Erlbaum.

Gothard, K. M., Skaggs, W. E., Moore, K. M., & McNaughton, B. L. (1994). Behavioral correlates of hippocampal CA1 cells in a landmark navigation task. *Society for Neuroscience Abstracts*, *20*, 1207.

Gould, S., & Lewontin, R. (1979). The spandrels of San Marco and the panglossian paradigm: A critique of the adaptationist programme. *Proceedings of the Royal Society, London, B*, *205*, 581–598.

Graf, P., & Schacter, D. L. (1987). Selective effects of interference on implicit and explicit memory for new associations. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *13*, 45–53.

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.

Greenough, W. T., Armstrong, K. E., Cummery, T. A., Hawry, N., Humphreys, A. G., Kleim, J., Swain, R. A., & Wang, X. (1994). Plasticity-related changes in synapse morphology. *Cellular and molecular mechanisms underlying higher neural functions*, Vol. 54 of *Dahlem Workshop Reports/Life Sciences* (pp. 211–220). New York: Wiley.

Grossberg, S. (1987). Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science*, *11*(1), 23–63.

Haist, F., Musen, G., & Squire, L. R. (1991). Intact priming of words and nonwords in amnesia. *Psychobiology*, *19*, 275–285.

Harris, E. W., Ganong, A. H., & Cotman, C. W. (1984). Long-term potentiation in the hippocampus involves activation of N-methyl-D-aspartate receptors. *Brain Research*, *323*, 132–137.

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

Hetherington, P. A., & Seidenberg, M. S. (1989). Is there 'catastrophic interference' in connectionist networks? *Proceedings of the 11th Annual Conference of the Cognitive Science Society* (pp. 26–33). Hillsdale, NJ: Erlbaum.

Hinton, G. E. (1981). Implementing semantic networks in parallel hardware. In G. E. Hinton, & J. A. Anderson (Eds.), *Parallel models of associative memory* (Chap. 6, pp. 161–187). Hillsdale, NJ: Erlbaum.

Hinton, G. E. (1989). Learning distributed representations of concepts. In R. G. M. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neurobiology* (Chap. 3, pp. 46–61). Oxford: Clarendon Press.

Hinton, G. E., & McClelland, J. L. (1988). Learning representations by recirculation. In D. Z. Anderson (Ed.), *Neural information processing systems* (pp. 358–366). New York: American Institute of Physics.

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed representations. In D. E. Rumelhart, & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition*, Vol. 1 (Chap. 3, pp. 77–109). Cambridge, MA: MIT Press.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*, 2554–2558.

Howe, M. L., & Courage, M. L. (1993). On resolving the enigma of infantile amnesia. *Psychological Bulletin*, *113*(2), 305–326.

Humphreys, M. S., Bain, J. D., & Pike, R. (1989). Different ways to cue a coherent memory system: A theory for episodic, semantic, and procedural tasks. *Psychological Review*, *96*, 208–233.

James, W. (1890). *Psychology (briefer course)*. New York: Holt.

Jarrard, L. E. (1989). On the use of ibotenic acid to lesion selectively different components of the hippocampal formation. *Journal of Neuroscience Methods*, *29*, 251–259.

Jarrard, L. E. (1993). On the role of the hippocampus in learning and memory in the rat. *Behavioral and Neural Biology*, *60*, 9–26.

Kaas, J. H. (1994). The reorganization of sensory and motor maps in adult mammals. In M. S. Gazzaniga (Ed.), *The cognitive neurosciences* (pp. 51–71). Cambridge, MA: MIT Press.

Keil, F. C. (1979). *Semantic and conceptual development: An ontological perspective*. Cambridge, MA: Harvard University Press.

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

Knapp, A., & Anderson, J. A. (1984). A signal averaging model for concept formation. *Journal of Experimental Psychology: Learning Memory and Cognition*, *10*, 617–637.

Knierim, J. J., Kudrimoti, H., & McNaughton, B. L. (in press). Hippocampal place fields, the internal compass, and the learning of landmark stability. *Journal of Neuroscience*.

Knowlton, B. J., Ramus, S. J., & Squire, L. R. (1992). Intact artificial grammar learning in amnesia: Dissociation of classification learning and explicit memory for specific instances. *Psychological Science*, *3*(3), 172–179.

Knowlton, B. J., & Squire, L. R. (1993). The learning of categories: Parallel brain systems for item memory and category knowledge. *Science*, *262*, 1747–1749.

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

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

Larson, J., & Lynch, G. (1986). Induction of synaptic potentiation in hippocampus by patterned stimulation involves two events. *Science*, *232*, 985–988.

Lee, K. S. (1983). Sustained modification of neuronal activity in the hippocampus and neocortex. In W. Seifert (Ed.), *Neurobiology of the hippocampus* (pp. 265–272). New York: Academic Press.

Leonard, B. J., McNaughton, B. L., & Barnes, C. (1987). Suppression of hippocampal synaptic plasticity during slow-wave sleep. *Brain Research*, *425*, 174–177.

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*, Vol. 23 of *The psychology of learning and motivation* (pp. 243–305). New York: Academic Press.

Levy, W. B., & Steward, O. (1979). Synapses as associative memory elements in the hippocampal formation. *Brain Research*, *175*, 233–245.

Linsker, R. (1986a). From basic network principles to neural architecture, i: Emergence of spatial-opponent cells. *Proceedings of the National Academy of Sciences, USA*, *83*, 7508–7512.

Linsker, R. (1986b). From basic network principles to neural architecture, ii: Emergence of orientation-selective cells. *Proceedings of the National Academy of Sciences, USA*, *83*, 8390–8394.

Linsker, R. (1986c). From basic network principles to neural architecture, iii: Emergence of orientation columns. *Proceedings of the National Academy of Sciences, USA*, *83*, 8779–8783.

Lynn, P. J. (1994). *System interaction in human memory and amnesia: Theoretical analysis and connectionist modeling*. PhD thesis, University of Colorado, Department of Computer Science, Boulder, Colorado.

MacKinnon, D., & Squire, L. R. (1989). Autobiographical memory in amnesia. *Psychobiology*, *17*, 247–256.

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

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

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

Maunsell, J. H. R., & Van Essen, D. C. (1983). The connections of the middle temporal visual area (mt) and their relation to a cortical hierarchy in the macaque monkey. *Journal of Neuroscience*, *3*, 2563–2586.

Mazzoni, P., Andersen, R. A., & Jordan, M. I. (1991). A more biologically plausible learning rule for neural networks. *Proceedings of the National Academy of Sciences USA*, *88*, 4433–4437.

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., McNaughton, B. L., & O'Reilly, R. C. (1993). Why do we have a special learning system in the hippocampus?, Abstract 580. *The Bulletin of the Psychonomic Society*, *31*, 404.

McClelland, J. L., McNaughton, B. L., O'Reilly, R. C., & Nadel, L. (1992). Complementary roles of hippocampus and neocortex in learning and memory, Abstract 508.7. *Society for Neuroscience Abstracts*, *18*, 1216.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General*, *114*, 159–188.

McClelland, J. L., & Rumelhart, D. E. (1988). *Explorations in parallel distributed processing: A handbook of models, programs, and exercises*. Boston, 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. 109–165). New York: Academic Press.

McNaughton, B. L. (1983). Comments in hippocampus symposium panel discussion. In W. Siefert (Ed.), *Neurobiology of the hippocampus* (pp. 609–610). New York: Academic Press.

McNaughton, B. L. (1989). The neurobiology of spatial computation and learning. In D. J. Stein (Ed.), *Lectures on complexity, Santa Fe Institute studies in the sciences of complexity* (pp. 389–437). Redwood, CA: Addison-Wesley.

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

McNaughton, B. L., Barnes, C. A., Rao, G., Baldwin, J., & Rasmussen, M. (1986). Long-term enhancement of hippocampal synaptic transmission and the acquisition of spatial information. *Journal of Neuroscience*, *6*, 563–571.

McNaughton, B. L., Douglas, R. M., & Goddard, G. V. (1978). Synaptic enhancement in facia dentata: Cooperativeity among coactive afferents. *Brain Research, 157*, 277–293.

McNaughton, B. L., Leonard, B., & Chen, L. (1989). Cortical-hippocampal interactions and cognitive mapping: A hypothesis based on reintegration of the parietal and inferotemporal pathways for visual processing. *Psychobiology, 17*, 230–235.

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.

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

Merzenich, M. M., Recanzone, G. H., Jenkins, W. M., & Grajski, K. A. (1990). Adaptive mechanisms in cortical networks underlying cortical contributions to learning and nondeclarative memory. *The Brain*, Vol. LV of *Cold Spring Harbor symposia on quantitative biology* (pp. 873–887). Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

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

Miller, K. D., & Stryker, M. P. (1990). Ocular dominance column formation: Mechanisms and models. In S. J. Hanson, & C. R. Olson (Eds.), *Connectionist modeling and brain function: The developing interface* (pp. 255–350). Cambridge, MA: MIT Press.

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 and Co.

Milner, B., Corkin, S., & Teuber, H.-L. (1968). Further analysis of the hippocampal amnesia syndrome: 14-year follow-up study of H.M. *Neuropsychologia, 6*, 215–234.

Milner, P. (1989). A cell assembly theory of hippocampal amnesia. *Neuropsychologia*, *27*, 23–30.

Minai, A. A., & Levy, W. B. (1993). Predicting complex behavior in sparse asymmetric networks. In S. J. Hanson, J. D. Cowan, & C. L. Giles (Eds.), *Advances in neural information processing systems 5* (pp. 556–563). San Mateo, CA: Morgan Kaufmann.

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 Press.

Monaghan, D. T., & Cotman, C. W. (1989). Regional variations in NMDA receptor properties. In J. C. Watkins, & G. L. Collingridge (Eds.), *The NMDA receptor.* (pp. 53–64). Oxford, UK: Oxford University Press.

Morris, R. G. M., Anderson, E., Lynch, G. S., & Baudry, M. (1986). Selective impairment of learning an blockade of long-term potentiation by an N-methyl-D-aspartate receptor antagonist, AP5. *Nature, 319*, 774–776.

Munro, P. W. (1986). State-dependent factors influencing neural plasticity: A partial account of the critical period. In J. L. McClelland, & D. E. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition*, Vol. 2 (Chap. 24, pp. 471–502). Cambridge, MA: MIT Press.

O'Keefe, J., & Conway, D. (1978). Hippocampal place units in the freely moving rat: Why they fire where they fire. *Experimental Brain Research*, *31*, 573–590.

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: Clarendon Press.

Olton, D., Becker, J., & Handelmann, G. E. (1979). Hippocampus, space, and memory. *Behavioral and Brain Science, 2*, 313–365.

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

Pavlides, C., & Winson, J. (1989). Influences of hippocampal place cell firing in the awake state on the activity of these cells during subsequent sleep episodes. *Journal of Neuroscience*, *9*(8), 2907–2918.

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

Qin, Y., Markus, E. J., McNaughton, B. L., & Barnes, C. A. (1994). Hippocampal place fields change with navigational context. *Society for Neuroscience Abstracts*, *20*, 1207.

Quillian, M. R. (1968). Semantic memory. In M. Minsky (Ed.), *Semantic information processing*. Cambridge, MA: MIT Press.

Quirk, G. J., Muller, R., & Kubie, J. L. (1990). The firing of hippocampal place cells in the dark depends on the rats recent experience. *Journal of Neuroscience, 10*, 2008–2017.

Ratcliff, R. (1990). Connectionist models of recognition memory: Constraints imposed by learning and forgetting functions. *Psychological Review*, *97*, 285–308.

Rawlins, J. N. P. (1985). Associations across time: The hippocampus as a temporary memory store. *Behavioral and Brain Sciences*, *8*, 479–496.

Ribot, T. (1882). *Diseases of memory*. New York: Apppleton-Century-Crofts.

Rips, L. J., Shoben, E. J., & Smith, E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior*, *12*, 1–20.

Rolls, E. (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* (Chap. 4, pp. 73–90). San Diego, CA: Academic Press.

Rozin, P. (1976). The evolution of intelligence and access to the cognitive unconscious. In A. Sprague, & A. N. Epstein (Eds.), *Progress in psychobiology and physiological psychology*, Vol. 6 (pp. 73–90). New York: Academic Press.

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

Rudy, J. W., & Sutherland, R. W. (1994). Configural association theory and the hippocampal formation: An appraisal and reconfiguration. *manuscript*.

Rumelhart, D. E. (1990). Brain style computation: Learning and generalization. In S. F. Zornetzer, J. L. Davis, & C. Lau (Eds.), *An introduction to neural and electronic networks* (Chap. 21, pp. 405–420). San Diego, CA: Academic Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart, & J. L. McClelland (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition*, Vol. 1 (pp. 318–362). Cambridge, MA: MIT Press.

Rumelhart, D. E., & Todd, P. M. (1993). Learning and connectionist representations. In D. E. Meyer, & S. Kornblum (Eds.), *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience* (pp. 3–30). Cambridge, MA: MIT Press.

Schacter, D. (1994). Priming and multiple memory systems: Perceptual mechanisms of implicit memory. In D. Schacter, & E. Tulving (Eds.), *Memory systems, 1994* (pp. 234–268). Cambridge, MA: MIT Press.

Schacter, D. L. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *13*, 501–518.

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.

Schacter, D. L., & Graf, P. (1989). Modality specificity of impliocit memory for new associations. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *15*, 3–21.

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

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, *3*, 207–217.

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

Shen, B., & McNaughton, B. L. (1994). Simulation of the spontaneous reactivation of experience-specific hippocampal cell assemblies during sleep. *Society for Neuroscience Abstracts*, *20*, 1206.

Shimamura, A. P., & Squire, L. R. (1987). A neuropsychological study of fact memory and source amnesia. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *13*, 464–473.

Shimamura, A. P., & Squire, L. R. (1989). Impaired priming of new associations in amnesia. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *15*, 721–728.

Singer, W., & Artola, A. (1994). Plasticity of the mature neocortex. *Cellular and molecular mechanisms underlying higher neural functions*, Vol. 54 of *Dahlem Workshop Reports/Life Sciences* (pp. 49–69). New York: Wiley.

Skaggs, W. E., McNaughton, B. L., Barnes, C. A., Moore, K. M., Duffield, C., Frank, L., & D'Monte, R. (1994). Sparse vs. distributed population coding in superficial and deep layers of rat neocortex. *Society for Neuroscience Abstracts*, *20*, 1206.

Sloman, S. A., & Rumelhart, D. E. (1992). Reducing interference in distributed memories through episodic gating. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *Essays in honor of W. K. Estes, Vol. 1* (pp. 227–248). Hillsdale, NJ: Erlbaum.

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

Squire, L. R., & Cohen, N. (1979). Memory and amnesia: Resistance to disruption develops for years after learning. *Behavioral and Neural Biology*, *25*, 115–125.

Squire, L. R., Cohen, N. J., & Nadel, L. (1984). The medial temporal region and memory consolidation: A new hypothesis. In H. Weingartner, & E. Parker (Eds.), *Memory consolidation* (pp. 185–210). Hillsdale, NJ: Erlbaum.

Squire, L. R., Haist, F., & Shimamura, A. P. (1989a). The neurology of memory: Quantitative assessment of retrograde amnesia in two groups of amnesic patients. *The Journal of Neuroscience*, *9*, 828–839.

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

Squire, L. R., & Slater, P. C. (1975). Forgetting in very long-term memory as assessed by an improved questionnaire technique. *Journal of Experimental Psychology: Human Learning and Memory*, *104*(1), 50–54.

Squire, L. R., Slater, P. C., & Chace, P. (1975). Retrograde amnesia: Temporal gradient in very long-term memory following electroconvulsive therapy. *Science*, *187*, 77–79.

Squire, L. R., Zola-Morgan, S., & Alvarez, P. (1994). Functional distinctions within the medial temporal memory system: What is the evidence? *Behavioral and Brain Sciences*, *17*, 495–496.

Stewart, C., & Reid, I. C. (1993). Electroconvulsive stimulation and synaptic plasticity in the rat. *Brain Research*, *620*, 139–141.

Sutherland, R. W., Arnold, K. A., & Rodriguez, A. R. (1987). Hippocampal damage produces temporally-graded retrograde amnesia in rats. *Society for Neuroscience Abstracts*, *13*, 1066.

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

Suzuki, W. A. (1994). What can neuroanatomy tell us about the functional components of the hippocampal memory system? *Behavioral and Brain Sciences*, *17*, 496–498.

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

Touretzky, D. S., & Geva, S. (1987). A distributed connectionist representation for concept structures. *The Ninth Annual Conference of the Cognitive Science Society* (pp. 155–164). The Cognitive Science Society, Hillsdale, NJ: Erlbaum.

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

Tulving, E. (1983). *Elements of episodic memory*. New York: Oxford University Press.

Warrington, E. K., & McCarthy, R. A. (1988). The fractionation of retrograde amnnesia. *Brain and Cognition*, *7*, 184–200.

Warrington, E. K., & Weiskrantz, L. (1978). Further analysis of the prior learning effect in amnesic patients. *Neuropsychologia*, *16*, 169–177.

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

Wickelgren, W. A. (1972). Trace resistance and the decay of long-term memory. *Journal of Mathematical Psychology*, *9*, 418–455.

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

Wigström, H., Gustaffson, B., & Huang, Y. Y. (1986). Mode of action of excitatory amino acid receplor antagonists on hippocampal long-lasting potentiation. *Neuroscience*, *17*, 1105–1115.

Wilson, M. A., & McNaughton, B. L. (1993). Dynamics of the hippocampal ensemble code for space. *Science*, *261*, 1055–1058.

Wilson, M. A., & McNaughton, B. L. (1994a). The preservation of temporal order in hippocampal memory reactivation during slow-wave sleep. *Society for Neuroscience Abstracts*, *20*, 1206.

Wilson, M. A., & McNaughton, B. L. (1994b). Reactivation of hipocampal ensemble memories during sleep. *Science*, *265*, 676–679.

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

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

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

Zola-Morgan, S., Squire, L. R., & Amaral, D. G. (1986). Human amnesia and the medial temporal region: Enduring memory impairment following bilateral lesion limited to field CA1 of the hippocampus. *Journal of Neuroscience*, *6*, 2950–2967.