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

**Article** *in* Psychological Review · August 1995

DOI: 10.1037/0033-295X.102.3.419 · Source: PubMed

**3 authors:**

James L Mcclelland
Stanford University
**358** PUBLICATIONS **76,825** CITATIONS
SEE PROFILE

Bruce L Mcnaughton
University of Lethbridge
**270** PUBLICATIONS **46,275** CITATIONS
SEE PROFILE

Randall C. O'Reilly
University of California, Davis
**179** PUBLICATIONS **20,093** CITATIONS
SEE PROFILE

# 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[1], Bruce L. McNaughton[2]
& Randall C. O'Reilly[1]

Carnegie Mellon University[1] & The University of Arizona[2]

Technical Report PDP.CNS.94.1
March 1994

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 supported by an ONR Fellowship. We thank Sue Becker, Loren Frank, Patrick Lynn, William Skaggs, and Matthew Wilson, and especially Lynn Nadel for discussions and/or useful comments on a draft of this paper. The computational argument developed in this article were presented in one abstract (McClelland, McNaughton, O'Reilly, & Nadel, 1992), and preliminary versions of the simulations of retrograde amnesia in rats and monkeys were previously presented in another (McClelland, McNaughton, & O'Reilly, 1993).

Requests for reprints should be addressed to the first author at the Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213.

# Abstract

The influence of prior experience on some forms of behavior and cognition is drastically affected by damage to the hippocampal system. However, if the hippocampal system is left intact both during the experience and for a period of time thereafter, subsequent damage can have much less or even no effect. Such findings suggest that memory traces change over time in a way that makes them less dependent on the hippocampal system. This process of change has often been called consolidation. Consolidation is a very gradual process; in humans, it appears to span up to 15 years. This article asks what consolidation is and why it occurs. We take as our point of departure the view that the initial memory trace that results from a relevant experience consists of changes to the strengths of the connections among neurons in the hippocampal system. Bidirectional connections between the neocortex and the hippocampus allow these initial traces to mediate the reinstatement of representations of events or experiences in the neocortex. Consolidation results from the cumulative effects of small, incremental changes to connections among neurons in the neocortex that occur each time such a representation is reinstated. This view leads to two key questions: 1) Why are plastic changes made initially in the hippocampus, if ultimately the substrate of a consolidated memory lies in the neocortex? 2) Why does consolidation span such an extended period of time?

Insights from connectionist network models of learning and memory provide one set of possible answers to these questions. These models consist of networks of simple processing units and weighted connections among the units, and they offer procedures for discovering what weights or values to use on the connections so that the network can capture the structure present in ensembles of events and experiences drawn from some domain. These connection weights then provide the basis for appropriate generalization to novel examples from the same domain. Crucially, the success of these procedures depends on interleaved learning: making only very small changes to the connection weights on each learning trial, so that the overall direction of weight change can be governed by the structure of the domain rather than the individual examples. The sequential acquisition of new data is incompatible with the gradual discovery of structure and can lead to *catastrophic interference* with what has previously been learned. In the light of these observations, we suggest that the neocortex may be optimized for the gradual discovery of the shared structure of events and experiences, and that the hippocampal system is there to provide a mechanism for rapid acquisition of new information without interference with previously discovered regularities. After this initial acquisition, the hippocampal system serves as teacher to the neocortex: That is, it allows for the reinstatement in the neocortex of representations of past events, so that they may be gradually acquired by the cortical system via interleaved learning. We equate this interleaved learning process with consolidation, and we suggest that it is necessarily slow so that new knowledge can be integrated effectively into the structured knowledge contained in the neocortical system.

## Contents

## Introduction

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 the phenomenon is temporally graded retrograde amnesia. Although this phenomenon has been controversial, a considerable body of research that we will describe in detail below supports the conclusion that the influence of prior experience on a wide range of behavioral and cognitive tasks is temporally circumscribed: In such cases, if the hippocampal system is damaged before or within a window of time after the initial experience, performance is impaired. 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.

The process underlying these gradual changes over time has often been called consolidation, but this 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 our discussion with a brief overview of the neuropsychology of memory, with particular emphasis on the temporally circumscribed role of the hippocampal system. We elaborate one possible account of the functional organization of the memory system that appears to be broadly consistent with the experimental evidence. We then describe some 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 are not detailed neural models; rather they illustrate at an abstract level what we take consolidation to be essentially about. We conclude with a brief comparison of our views to those of others who have theorized about the role of the hippocampal system in learning and memory. We note that our approach differs markedly from many other approaches, in that it treats gradual transfer to the neocortex as reflecting a principled aspect of the design of the mammalian memory system rather than just an artifact or an unmotivated performance limitation.

## Role of the Hippocampal System in Learning and Memory

We use the phrase *the hippocampal system* to refer to a system that appears to be specialized for the rapid formation of certain kinds of new memories. It consists of the *hippocampus proper* together with adjacent areas making up the *parahippocampal region*, located in the medial temporal lobes (this terminology comes from Eichenbaum, Otto, & Cohen, in press). We contrast the hippocampal system with what we call *the neocortical system* or just *the neocortex*, by which we mean those parts of the brain thought to be required for all sorts of information processing tasks, including perception, action, language, attention, etc. This system may include the basal ganglia, cerebellum, amygdala, and several other brain regions as well as the neocortex proper. The functional boundary between the hippocampal and neocortical systems is not clearly delineated at this time and there may well be some overlap.

The literature on the role of the hippocampal system in learning and memory is quite vast. Here we summarize what we believe are the main points:

(1) Hippocampal lesions can produce a profound deficit in new learning, while leaving memory performance based on old material, acquired well before the lesion, at least clinically 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 hippocampus and related structures located in the mesial portions of the temporal lobe. 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 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. It necessarily oversimplifies somewhat to give a succinct summary of the huge literature on this point, but the following characterization appears to provide a useful first approximation: The hippocampal system appears to be essential for the rapid formation of new arbitrary associations or conjunctions of the various contents or elements of specific events and experiences, in a form sufficient to sustain an explicit retrieval of the contents of the association, so that it can be explicitly recognized and verbally described (in humans), or flexibly used to govern subsequent behavior. Squire (1992) proposes the terms *explicit* and *declarative*, and these appear to be useful at least as shorthand characterizations; Squire's notion is that explicit memory depends on the rapid binding together of all of the elements that make up the episode or event. Included in the category of explicit memories are what Tulving (1983) termed *episodic memories* – 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. A paradigm example of a task that depends on this form of memory is paired-associate learning with arbitrary word pairs, in which the subject sees a list of word-pairs in a study session, and is later asked to recall the second member of each pair, when given the first as a cue. Prior associations to the cue word are unhelpful in this task; to be successful, the subject must recall the word that occurred previously with the cue word in the list study context, and since the earliest tests on HM is has been clear that hippocampal lesions produce profound impairments in learning arbitrary paired associates (Scoville & Milner, 1957). In the animal literature, the hippocampus has been described as crucial for learning to make appropriate responses that depend on combinations of cues or *cue configurations* (Sutherland & Rudy, 1989), while Cohen and Eichenbaum (1993) have emphasized the importance of flexible access to memory traces for use in governing behavior, a characteristic that may be closely related to explicit recall 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. However, both in humans and animals, large effects of hippocampal system lesions are often obtained in tasks that make no obvious spatial demands. Thus, we have taken the view that tasks that tap memory for location are simply good examples of tasks that require the use of the form of learning that occurs in the hippocampal system, since the sense of place depends on complex combinations of cues, and place is a crucial aspect of context. In recent years, it has become

apparent that some fractionation of the effect of lesions to the hippocampal system may be possible. Lesions to the hippocampus proper, sparing the structures in the parahippocampal region, can produce quantitatively smaller deficits, or deficits that show up only on certain tasks and or at long delay (see Squire, 1992; Eichenbaum et al., in press for reviews). We discuss this issue further in the *General Discussion*.

(3) Some kinds of learning appear to be completely unaffected by hippocampal lesions. Squire (1992) characterizes these forms of memory as *non-declarative* or *implicit*, emphasizing that they influence behavior without themselves giving rise to an explicit, declarative memory for the events that led to these influences (Schacter, 1987). 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). A second form of spared learning is 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 (Graff, 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 (facilitation can be measured in terms of percent correct at a given level of stimulus degradation, or reaction time to make a response), as long as care is taken to avoid the possibility that explicit recall is used to aid performance. In the animal learning literature, spared learning is often exhibited in tasks requiring a simple association between a simple stimulus such as a tone or a light and a response, or when adequate performance can be achieved through the exact repetition of a behavior previously acquired in a given situation (See Cohen & Eichenbaum, 1993; Eichenbaum et al., in press for full discussions).

(4) Insults to the hippocampal system, including lesions or bilateral electro-convulsive shock treatments, appear to give rise to a temporally graded retrograde amnesia for material acquired in the period of time preceding the lesion. Although these findings have been the subject of controversy (Warrington & Weiskrantz, 1978; Warrington & McCarthy, 1988), we believe they are substantial enough to be taken seriously, and they play a major role in the theory of hippocampal function 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.

- In some severe cases, the retrograde gradient can extend over periods of 15 years or more (Squire, Haist, & Shimamura, 1989; MacKinnon & Squire, 1989).

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 and mice (Winocur, 1990; Kim & Fanselow, 1992) have produced retrograde gradients covering a period of days or weeks; thus in Winocur (1990), hippocampal rats showed normal memory for material they had acquired as little as 10 days prior to surgery. They were impaired relative to controls only when the lesion occurred 0 to 5 days after the learning event. Primate experiments (Zola-Morgan & Squire, 1990) show a severe impairment relative to controls for memory acquired 2 or 4 weeks prior to surgery, but no reliable differences between normals and controls for older memories.

A key aspect of the retrograde amnesia findings that is now beginning to emerge is the finding that tasks that show little dependence on the hippocampal system for initial learning also show little or no retrograde gradient. A clear example of this comes from the study of Kim and Fanselow (1992) in which they showed that retention of a simple association between a tone and shock was as strong in rats with immediate post-exposure lesions as in intact animals and animals with delayed lesions.

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 an alternative interpretation that is consistent with some of the data. According to this alternative account, memories are initially stored in two forms, one that is hippocampal-system dependent and one that is not. The account further assumes that the hippocampal-dependent form of memory decays more rapidly than the non-hippocampal dependent form. On this account, there is no alteration of the form of memory over time, there is merely decay; but nevertheless, as memory performance becomes progressively worse with time for both hippocampals and controls, 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 at test is better when there is a longer delay between study and test, supporting a real change in the form 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 TV shows that aired for a single season, and Squire and Slater (1975) went to great lengths to demonstrate that acquisition of knowledge of the TV shows depended primarily on exposure to the shows during the year they were aired, and not on later exposure in subsequent years. The interpretation of the human data is clouded by the fact that in this case the amnesia was induced by treatment of severe depression, and it is difficult to rule out the possibility

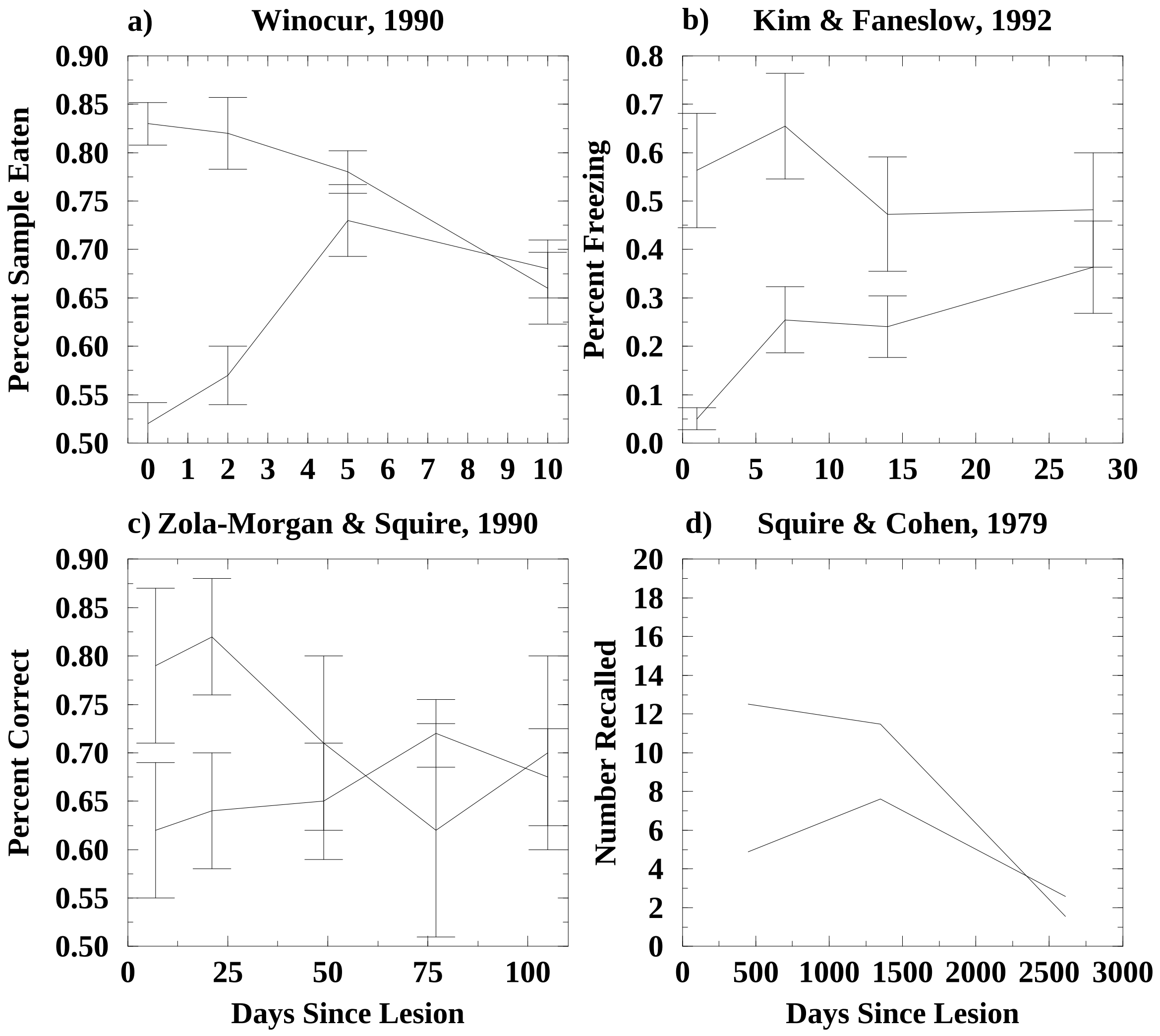

Figure 1: Panels (a) - (c) show behavioral responses of animals lesioned different numbers of days after exposure to relevant experiences, while panel (d) shows a comparable pattern from an experiment with human subjects. (a) Fear behavior shown by rats exposed to a contingency between tone and foot shock in a novel environment (Kim & Fanselow, 1992). (b) Food preferences exhibited by rats exposed to a conspecific (Winocur, 1990). (c) Choices of reinforced objects by monkeys exposed to 14 training trials with each of 20 object pairs (Zola-Morgan & Squire, 1990). Panel (d) shows recall by depressed human subjects of details of television shows aired different numbers of years prior to the time of test, either before of after ECT (Squire & Cohen, 1979). Note that we have translated years into days to allow comparison with the results from the animal studies.

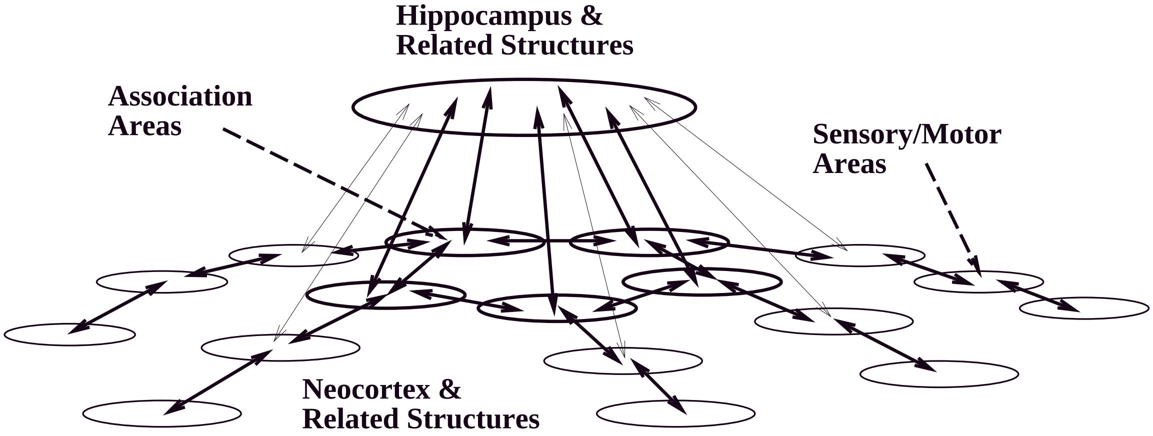

Figure 2: Conceptual sketch of the neocortical system and its connections with the hippocampal system.

that the depression in the years just prior to treatment affected initial storage of memories from this period. But no such interpretation is available for the findings from the animal studies, which are very clear in two out of the three 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 a brief statement of one possible account of the mechanisms of learning and memory in the mammalian brain, and of the expression of learning and memory in performance. The account is consistent with the data summarized above. Our proposals have commonalities with many aspects of the proposals offered in several other synthetic treatments ranging from Marr (1971) to Squire, Cohen, and Nadel (1984), McNaughton and Morris (1987), Milner (1989), Squire (1992) and Treves and Rolls (in press), though there are certainly differences between our proposals and each of these existing treatments. Discussion of important aspects of these similarities and differences may be found in the *General Discussion*.

(1) The neocortical system consists of a number of neural populations or brain areas widely distributed over the neocortex and other structures. These brain areas are strongly interconnected, usually with bidirectional projections, so that activity in one region can give rise to activity in many other regions. 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. The cortical representation of an experience or episode of information processing consists of a widely distributed pattern of activity. Of course, different experiences vary in the extent to which particular modalities and effector systems are involved. A highly schematic rendering designed to evoke the general organization is shown in Figure 2.

(2) Performance of cognitive tasks depends on the elicitation of patterns of activation

over the population of neurons in one region or regions by other patterns of activation over the same or different regions. In reading a printed word, for example, the pattern produced in visual cortex may elicit activity that corresponds to the sound of the word in some other region. In a free association task, a pattern representing a stimulus word elicits another pattern representing the response. In retrieving an arbitrary associate that was paired with a target word in an paired-associate learning task, the stimulus pattern must specify not only the stimulus word but also some information about the encoding context, but otherwise the basic principle remains the same: task performance occurs through the elicitation of one pattern of activation in response to another that serves as a cue. One form of associative elicitation, in which some subset of a larger pattern serves as cue for the elicitation of the entire pattern, is called pattern completion. Systems that perform pattern completion are basically content addressable memories; that is, any aspect of the content of the memory can serve to retrieve the entire memory. Both memory for the contents of specific episodes and memory for semantic information in the form of propositions about objects and events appear to operate in this way, and so we view them as examples of pattern completion processes.

(3) Ultimately, the knowledge that underlies all cognitive capacities, including acquired skills such as reading, acquired factual knowledge such as the names of friends and relatives, knowledge of word associations, semantic and encyclopedic knowledge about the nature and properties of objects, and even knowledge about specific remote episodes involving ourselves as participants, is stored in connections among neurons in the neocortical system. The knowledge is embedded in the connections among the very neural populations that carry out the tasks that use the information (i.e., that underlie the pattern elicitation processes involved).

(4) On every occasion of information processing, small adjustments to the connections among the neurons involved in the processing are made. The adjustments are widely distributed across all of the relevant connections, but are very small in magnitude, and so have relatively subtle effects; they are not sufficient to create a novel association between one item and another all at once, but they are sufficient to facilitate subsequent processing in cases where pre-existing connections provide a pre-existing substrate for performance.

(5) Over the course of many repetitions the changes in the cortical system will accumulate. When these capture the same content, they can provide the basis for correct performance in semantic, encyclopedic, and episodic memory tasks. When they reflect different examples of some sort of structured relationship between inputs and outputs, they provide the basis of acquired cognitive skills.

(6) When we store the contents of a specific episode or event in one or a few presentations (within, let us say, the standard hour of the typical memory experiment), in a form sufficient to allow later explicit recall or flexible use of the contents of the episode or event, we do so primarily via changes in the strengths of connections among neurons in the hippocampal system. Although each experience or episode does produce the slight changes already mentioned within the neocortical system, our suggestion is that these are not sufficient to subserve new arbitrary associations after just one or a few presentations. Information is carried between the hippocampus and the neocortex via bi-directional, relatively implastic pathways (these pathways may be multi-synaptic). During learning, the pattern representing some to-be-learned association as represented in the cortex gives rise via these connections

to a corresponding pattern of activation in the hippocampal system. Long-term potentiation of synapses within the hippocampal system then makes this pattern an attractor — that is, a pattern toward which other neighboring patterns will converge. During recall, if a part of the pattern representing the episode arises again in neocortex, this generates an input to the hippocampal system. If the input is sufficiently close to the stored attractor pattern, it will lead the hippocampal system to settle into the attractor. The return pathways from the hippocampal system to the neocortical system then complete the reinstatement of the corresponding neocortical pattern, and overt performance in memory tasks is then based on further processes that ensue, such as, for example, overt report of some aspect of the pattern reinstated in the neocortex.

(7) Hippocampus-based reinstatement may occur in task situations where the memory trace is needed, or in off-line situations, possibly including active rehearsal, reminiscence, and other inactive states including sleep (these situations are considered in more detail later). In all these cases, we assume that the reinstatement of the neocortical pattern provides a training trial for the neocortical system. To the extent that the hippocampus participates in this reinstatement process, it can be viewed not just as a memory store but as the teacher of the neocortical system.

It should be relatively easy to see how these proposals provide at least a descriptive account for the pattern of deficits found following a lesion to the hippocampal system. The profound deficit in the ability to form new arbitrary associations from a few exposures would arise from the fact that these would have been stored in the hippocampal system; the small changes that would occur in the neocortical system would be insufficient for adequate performance. The spared acquisition of skills would arise because these are gradually acquired, directly in the connections among the relevant neural populations in the cortical system itself. Spared priming would reflect the subtle effects of the small increments to neocortical connections that occur in each processing episode, and the temporally extended and graded nature of retrograde amnesia would reflect the fact that information initially stored in the hippocampal system would become incorporated into the cortical system only very gradually, due to the small size of the changes made on each reinstatement. The ability of even very profound amnesics to gradually acquire very common and often-repeated new material (Milner et al., 1968; Glisky, Schacter, & Tulving, 1986; Glisky & Schacter, 1986) would likewise reflect this slow accumulation of changes in the cortical system after the onset of amnesia.

Our proposals concerning the basis of performance in semantic, encyclopedic, and episodic memory tasks may require some further elaboration. As we define these terms here, semantic memory tasks are simply those that require the use of general factual information such as the characteristic and defining features of members of categories and the meanings of words; encyclopedic tasks are ones that require the use of specific factual information about historical events and specific persons, places, ideas, etc; and episodic memory tasks are those that require the use of information about a specific event or previous experience that happened to the subject. In our view, there is no special distinction between the kinds of knowledge underlying performance in such tasks, as the knowledge underlying all three types of task is eventually stored within the neocortical system via the gradual accumulation of the small connection changes that result from repeated activation of overlapping patterns in the neocortical system. To see how we view this as occurring in practice, let us consider the specific episode or event in which one first encounters some important fact: The fact, for example, that Neil Armstrong uttered the words "That's one small step for {a} man, one giant leap for mankind" when he first set foot on the moon. When this factual information is encountered it is encountered in a particular context. Let us suppose for the sake of the example that our subject (say a male college student at the time) encounters this information by watching Armstrong live on TV as he sets foot on the moon, during a family reunion celebrating his grandfather's 70th birthday. If this same factual information is reinstated repeatedly, the accumulated changes to neocortical connections arising from this would, in our proposal, eventually come to preserve the common aspects of these events and experiences. The result may be the accumulation of connection changes that would allow the individual to perform correctly in an encyclopedic memory task, for example a task in which the individual is asked to recall what Armstrong actually said. If these reinstatements occur in many different contexts, the connection changes relevant to the common content would accumulate but those relevant to associations between this content and the different contexts in which it occurred would not. We envision much the same kind of process for 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 (McClelland & Rumelhart, 1985 explicitly discuss how semantic memories and category learning can arise from the gradual accumulation of small changes resulting from individual events and experiences; a similar model of category learning has also been described by Knapp & Anderson, 1984). Finally, let us consider an episodic memory task; for example, recall of where we were and who we were with when we actually heard Armstrong utter his famous words. 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. Thus if the fact that we had watched the moon walk during the 70th birthday celebration were repeatedly reinstated during subsequent family reunions, the connection changes that would arise as a result of these repeated reinstatements would come to encode not only information about what Armstrong said but also the context in which he said it. Thus on our view there is no real distinction between the accumulated neocortical knowledge that underlies performance in semantic, encyclopedic, and remote episodic memory tasks.

The account we have given is intended to serve as a putative description of the role of the hippocampal system in learning and memory. Though it shares a great deal with other approaches, our account still embodies many unproven propositions, and so it might be viewed as a theory of memory in one sense. However, the account is not meant so much as a theory as a starting place for theoretical discussion. The account appears to be compatible enough with the data on hippocampal amnesia to treat it as provisionally correct. Supposing for the sake of argument that it is correct, we can then ask, why is it that the system is organized in this particular way?

## Key Questions about the Organization of Memory

If something like our provisional account is correct, two key functional questions arise:

- Why do we need a hippocampal system, if ultimately performance in 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 simulation models that adhere to many aspects of the account of the mammalian memory system given above, but which do not incorporate a special hippocampus-like system for rapid acquisition of episodic information. Such systems may consist of several modules and pathways interconnecting the modules, but are monolithic in the sense that there is no separate system for rapid acquisition of the contents of individual episodes and events.

### *Discovery of Shared Structure through Interleaved Learning*

The first, and perhaps the 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 the adjustments made to connection weights on each learning trial are sufficiently 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 does capture approximately the constraints that may be operative in the discovery of conceptual structure from linguistic input, 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 appearance of the concepts themselves but in the relations they enter into with other concepts.

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 (note that we refer here 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 3). Such networks are not to be confused with connectionist or neural 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 concept of egret to the concept of bird by an *isa* link. Then egret 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). They are cumbersome to use when they contain a large amount of information, particularly if we note that every concept is a constituent of multiple intersecting hierarchies (Fahlman, 1981). Also, such representational schemes lead to grave difficulties in determining 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.

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 internal representations to concepts that capture their conceptual similarity relations to each other. 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,

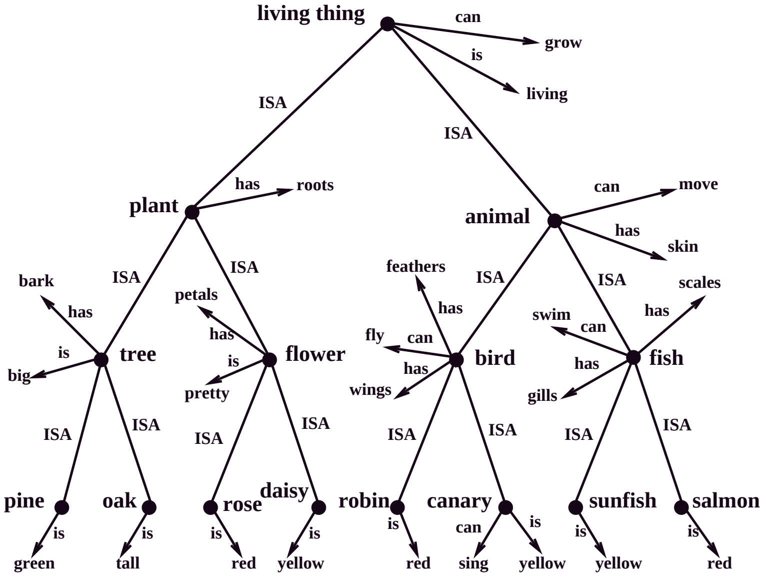

Figure 3: A semantic network of the type formerly used in models of the organization of knowledge in memory. After Rumelhart and Todd (1993).

plants, animals, trees, oaks, pines, flowers, roses, daisies, animals, birds, canaries, robins, fish, salmon, and sunfish. The network is shown in Figure 4. 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

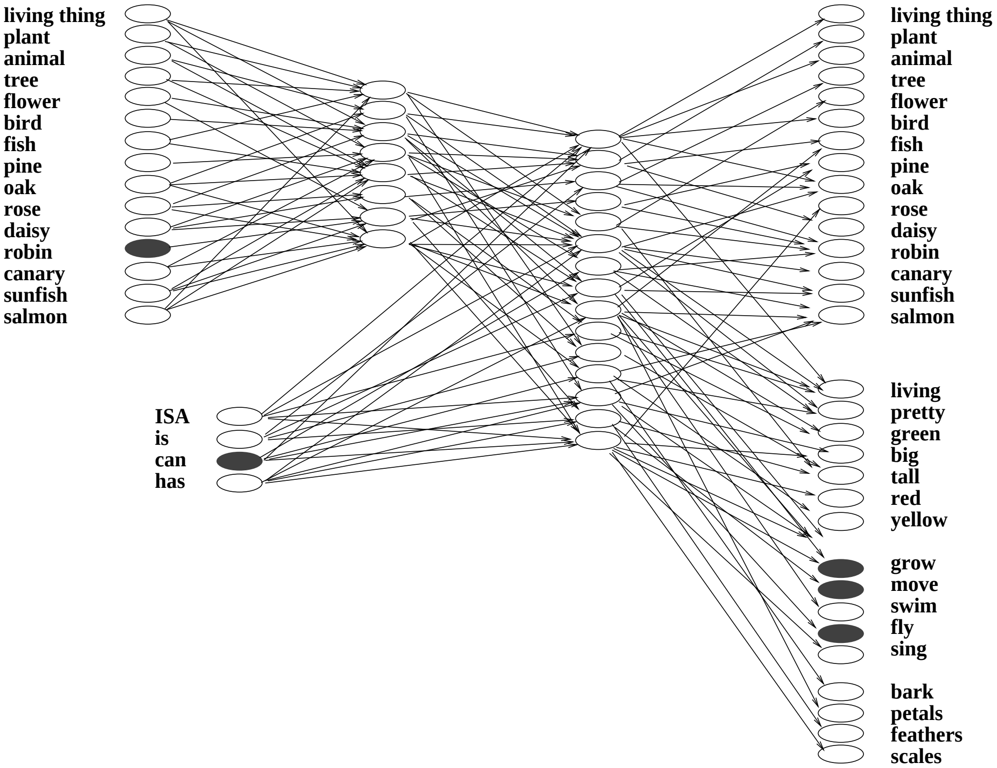

Figure 4: The connectionist network used by Rumelhart to learn propositions about the concepts shown in the Figure 3. The entire set of units used in the actual network is shown. Inputs are presented on the left, and activation propagates from left to right. Where connections are indicated, every unit 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.

by a learning rate constant $\epsilon$ that is set to a small value, so that only small changes are made on any given training trial. Thus, responses are learned slowly. 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 [1].

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.

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. The representations at different points in training are shown in Figure 5. 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 6, reflects the similarity structure among these patterns, as determined by a simple 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 more similar to *canary* than it is to *pine*. At later points, however, the 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

[1] All of the simulations reported here were done using the **bp** program of McClelland and Rumelhart (1988). 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 $\epsilon$ was set to 0.1. Targets for learning were .95 for units that should be "on" and .05 for units that should be "off".

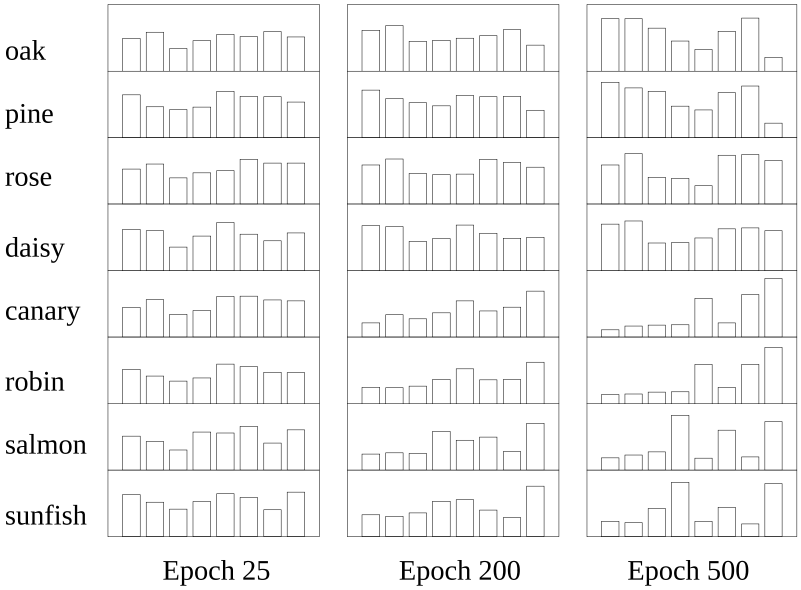

Figure 5: Representations discovered in our replication of Rumelhart's learning experiment, using the network shown in Figure 4. The figure shows the activation of each of the eight concept representation units for each of the eight specific concepts. The height of each vertical bar indicates the activation of the 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. After 500 epochs, the further differentiation of the plants into tress and flowers and of the animals into fish and birds is apparent.

any of the plants. Examination of the clustering of the representations at Epoch 200 shows that the network first learns 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 systematic with respect to subtype.

The similarity structure shown in Figure 6 — 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 structure in the inputs, but because of similarity structure in 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 the other relation terms. Therefore the network will respond appropriately when the other relation

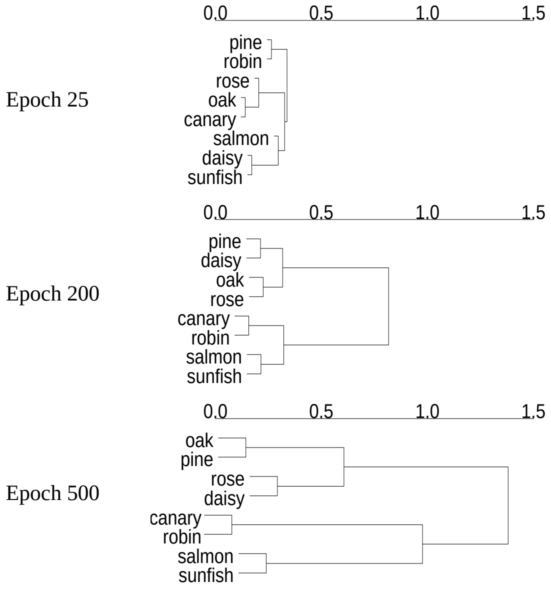

Figure 6: Similarity structure discovered in our replication of Rumelhart's learning experiment, using the representations shown in Figure 5. Initially, the patterns are all quite similar, and the weak similarity structure that exists is random. The concepts become progressively differentiated as learning progresses.

terms are paired with sparrow, even though it has never been trained on these cases. (In fact the network correctly activates all those outputs on which robin 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 process of cognitive development that unfolds gradually over many years. The order of acquisition of conceptual distinctions in such systems, beginning with coarser distinctions such as the one studied here between plants and animals and proceeding to the 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 finds of things; not so much in terms of the specific facts they knew about each, but in terms of the kinds of things that could be *predicated*, or attributed, to different kinds of things. 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 animals, but not to plants or inanimate objects. To assess children's knowledge of these matters, Keil asked children to indicate whether it was "silly" 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 Keil interpreted this as evidence that the child felt that the kind of property could be predicated of the thing in question. Based on children's judgments, Keil constructed what he called *pred-*

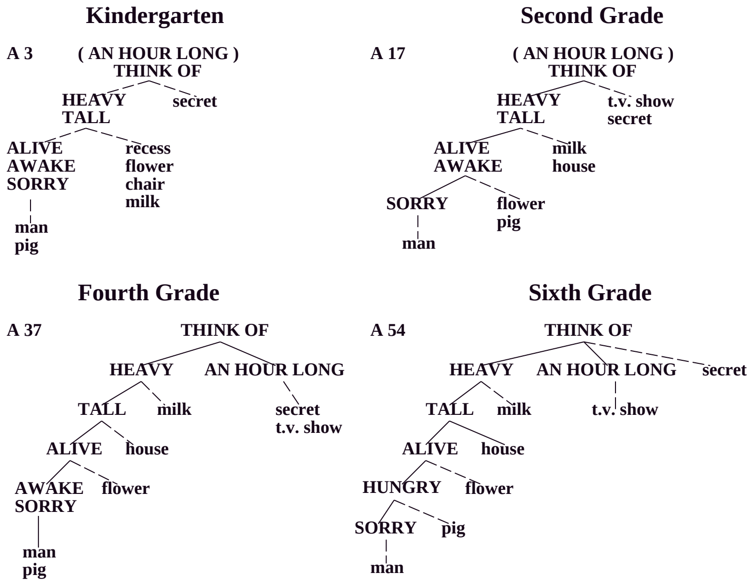

Figure 7: Predicability trees empirically derived by Keil (1979). The trees indicate the types of predicates children of different ages are willing to accept as applicable to different types of concepts at different ages. The trees were derived by asking children to whether they thought statements like "This chair is an hour long" were "silly". See text for further discussion.

*icability trees* for individual children. Four such trees, from children in different age groups, are shown in Figure 7. 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.

Keil's (1979) developmental findings mirror the progressive differentiation of concepts that we have seen in the Rumelhart 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, in press).

## *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 (McCloskey & Cohen, 1989; Ratcliff, 1990) as is required, for example, in paired-associate learning experiments (e.g., Barnes & Underwood, 1959). When used in such tasks, monolithic connectionist systems exhibit a phenomenon McCloskey and Cohen (1989) termed *catastrophic interference*.

McCloskey and Cohen used a connectionist network slightly simpler than the one used by Rumelhart to study this phenomenon. They were particularly interested in a paradigm called

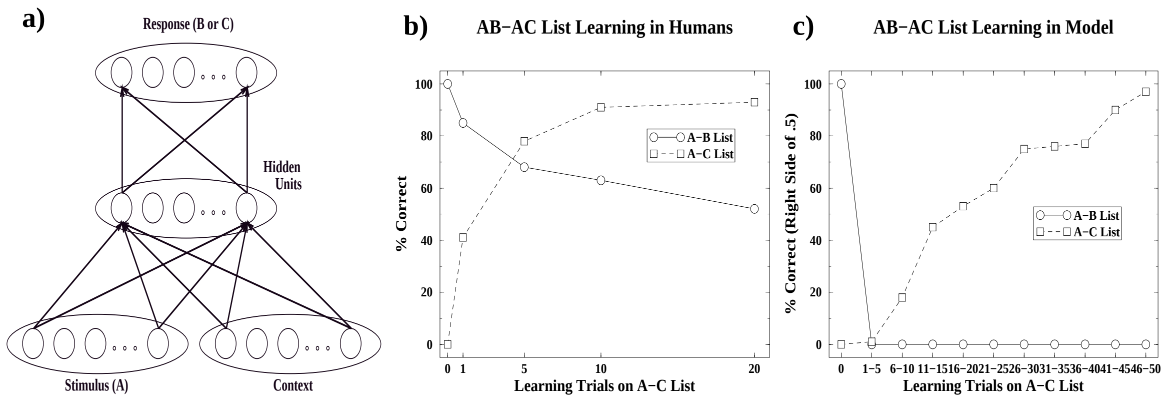

Figure 8: a) The network used by McCloskey and Cohen to demonstrate catastrophic interference, together with b) experimental data showing mild interference in humans in the AB-AC paradigm (Barnes & Underwood, 1959), and c) simulation results demonstrating catastrophic interference. Redrawn from McCloskey and Cohen (1989).

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. After 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 8a). 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 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 8c, and contrasted with typical human results in Figure 8b. 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 multi-layer networks. In fact, several investigators have demonstrated ways of reducing the magnitude of interference in simple networks like the one used by McCloskey and Cohen (McRae & Hetherington, 1993; Kortge, 1993). There are many techniques that can be used to reduce the magnitude of the crosstalk between patterns learned sequentially (Sloman & Rumelhart, 1992; Scalettar & Zee, 1986). 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. Reducing overlap does produce networks that avoid catastrophic interference — but the cost of reducing overlap is 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*

Before we address these issues explicitly, it is important to 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. But this issue can be explored in the Rumelhart 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 in new information is added gradually, interleaving it with ongoing exposure to examples of the domain of knowledge with which it interrelates.

We illustrate these points by examining what happens if we teach Rumelhart's network some new facts that are somewhat inconsistent with the existing knowledge in the system: the knowledge that penguins are birds, together with the knowledge that they can swim, but 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, we can train the network on the new material about penguins fairly quickly (Figure 9a). 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. What we see is a deleterious effect the new learning on the network's performance with other concepts – particularly other birds. What happens is that as the network learns that the penguin is a bird that can swim but not fly, it

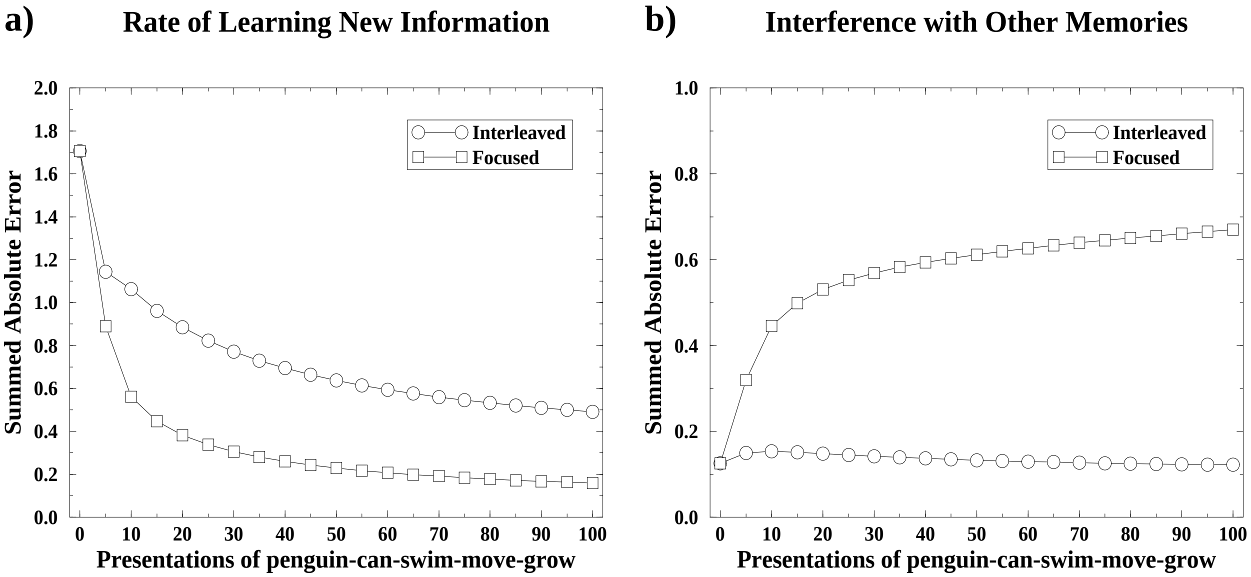

Figure 9: Effects of focused and interleaved learning on the acquisition of new knowledge and on interference with existing knowledge. Simulations were carried out using Rumelhart's network, using the connection weights resulting from the initial 500 epochs of training with the base corpus.

comes to treat all birds as having the same characteristics. Thus, focused learning results in very substantial interference with the networks existing store of knowledge (Figure 9b). With interleaved learning (also shown in Figure 9), 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, that 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 in the way of interference. Eventually, with enough practice, the network can in fact learn to activate strongly the correct output for the input *penguin-can*, without producing more than a very small ripple of interference with what it already knows about other related concepts. This is because the interleaved learning allows the network to carve out a place for the penguin in its structured system of knowledge, adjusting its representation of other similar concepts and adjusting its connection weights to incorporate the penguin into its structured knowledge system.

The observation that interleaved learning allows new knowledge to be gradually incorporated into a structured system of knowledge 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 et al. (1984, p. 205):

> ...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.

## Three Principles of Connectionist Learning

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

1. 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.

2. Attempts to learn new information rapidly in a network that has previously learned a subset of some domain leads to catastrophic interference.

3. 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 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?

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 other experiences, so that eventually the new knowledge may be incorporated into the structured system of knowledge 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 considered the discovery of shared structure through interleaved learning and catastrophic interference in focused learning through a very specific example. 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.

### *Quasi-Regularity, Partial Arbitrariness, and Memory for Facts and Experiences*

It should be noted that connectionist networks that are capable of extracting shared structure can also learn to accommodate arbitrary material, and material that has some arbitrary aspects. In this context it is worth considering as an example the learning of exception words in a system that learns 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 adult English speaking college students (Plaut & McClelland, 1993).

We think of the domains encompassed by semantic, episodic, and encyclopedic knowledge as quasi-regular, and we view facts and experiences as 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, etc. 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. Integration of the contents of these experiences into structured knowledge systems is therefore a crucial aspect of semantic, episodic, and encyclopedic memory.

### *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:

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

Here $w_{ij}$ is the weight to unit $i$ from unit $j$, and $\Delta w_{ij}$ represents the change in this weight. The variables $a_i$ and $a_j$ represent the activations of units $i$ and $j$, and $\epsilon$ 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 $\epsilon$. The only meaningful values of $\epsilon$ are positive real numbers $\leq 1$. When $\epsilon$ 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 $\epsilon$, 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. We need 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 in this case there are at least two reasons why the discovery of structure requires the use of a small learning rate.

## *Discovery of Structure in Biologically Realistic Systems*

Let us now consider the process of discovering structure at 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, 1986c, 1986a, 1986b; Miller, Keller, & Stryker, 1989, Miller & Stryker, 1990; Kohonen, 1990). For example, Linsker (1986c) 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 $\kappa$ 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 much 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. 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 appropriate representation of one part of the 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 bidirectional communication among the different parts of the system.

The major breakthrough in connectionist learning was the discovery of procedures, more powerful that 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 appears to remain 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 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 structure in multi-layer networks.

One solution to this problem comes from the idea that learning in multilayer systems might exploit the reciprocity of connections 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, Hinton, & Sejnowski, 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 degeneration 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 training time 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. It is, of course, our contention that it is precisely to solve the problem of allowing 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.

Hippocampally-mediated reinstatement can be divided into two types of cases: task-relevant and task-irrelevant. By task-relevant reinstatement, we mean reinstatement that occurs when the memory is actually being put to use in some situation where it is necessary for performance. Obviously this type of hippocampus-dependent use of memory does occur; but if our theory is correct and consolidation occurs only through reinstatement of memory traces in the neocortex, there must also be task-irrelevant reinstatement, since, in most of the animal experiments on consolidation, the process occurs during periods when the animals have no exposure to the task or even the locations in the environment in which the memory trace was originally formed.

A number of possible mechanisms for task-irrelevant reinstatement can be imagined. As pointed out by Marr (1971) it is most likely that such reactivation would occur during periods when the hippocampus is not actively engaged in processing external inputs, such as during quiet wakefulness or sleep. In humans, there is clearly the phenomenon of reminiscence, which can be considered as reinstatement in a task-irrelevant context while thinking about the task situation. It seems possible that reminiscence can occur in other animals as well, though it may be difficult to obtain data that would specifically document explicit reminiscence in nonhumans. In both rodents and primates, during periods of quiet wakefulness and slow-wave sleep, hippocampal electrical activity is characterized by a unique pattern called sharp waves or ripples (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, one of the parahippocampal regions. Thus, patterns stored in the hippocampus might complete themselves during hippocampal sharp waves, thereby providing an opportunity for reinstatement in the neocortex. In support of this idea, 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 (1993) have found that the cross-correlation structure that arises in a large population (50-100) of simultaneously recorded CA1 neurons during exploration of a novel 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, correlated activity associated with such traces — are indeed reactivated in the rat hippocampus during "off-line" periods.

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. In our view, 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. It might be supposed that such arbitrary material should not be consolidated at all, since it does not really contribute to the structure such systems are intended to extract. However, as we discussed previously in the section on quasi-regular domains, even experiences that have arbitrary elements generally share some structure with many other experiences, and consolidation of such material should contribute to its exploitation.

For complete consolidation of the contents of a partially arbitrary association, the neocortical system will need to find a set of connection weights that accommodate both the common and the idiosyncratic aspects. Those 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 & McClelland, 1993). 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, if it persists for long enough in the hippocampal system.

## *Modeling Temporally Graded Retrograde Amnesia*

To illustrate our conception of the consolidation process, we undertake in this section to provide simulations of several experiments in the growing literature on retrograde amnesia. We begin by considering recent studies in the animal literature, in which the physiological manipulation is a bilateral lesion to some or all of the hippocampal system at some time after exposure to some learning experience. The focus here is on the consolidation of arbitrary material. Thus, the simulations do not address the gradual discovery of structure, but only the interplay between hippocampal decay and consolidation of arbitrary, idiosyncratic associations.

In the simulations that follow, we do not actually attempt to simulate the formation of memories in the hippocampal system. Rather, we treat the hippocampus as a black box, and show that an account can 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 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, the hippocampal memory traces degrade. This could occur either as a result of passive decay of the relevant enhanced connections or as a result of interference. Third, we assume that the probability of hippocampally mediated reinstatement in the neocortex decreases with the quality 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 that memories of only transitory relevance. In any case, in the remainder of this section we ignore the effects of self-reinforcement for simplicity, noting that such effects, if they exist, would affect 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. The neural 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

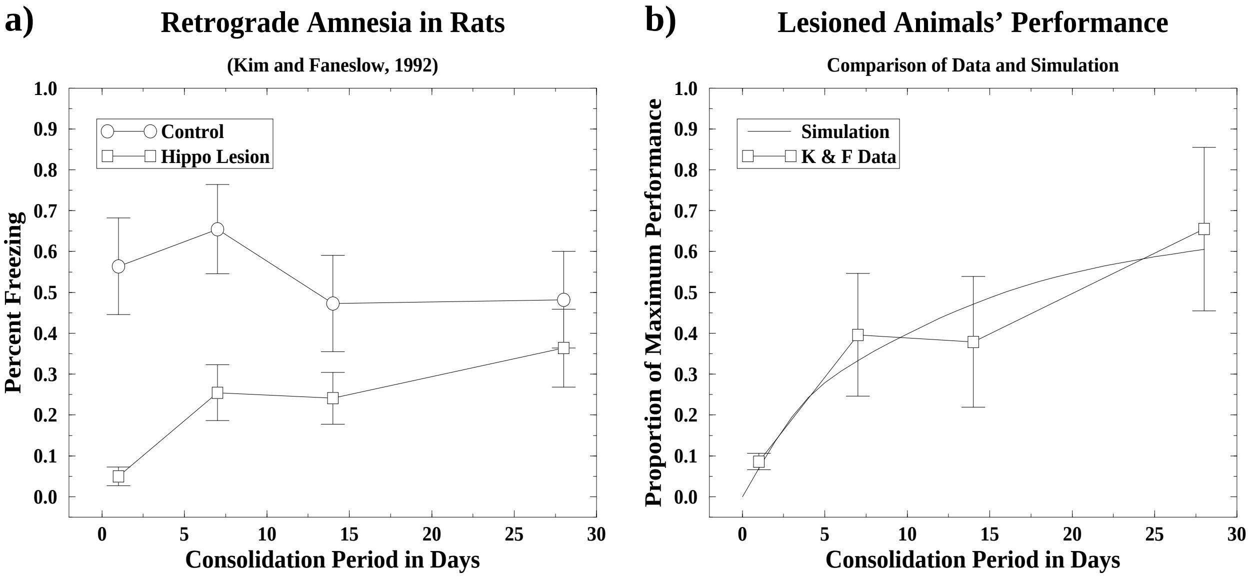

Figure 10: Consolidation as observed in the experiment of Kim and Fanselow (1992). a) Data from experimental and control groups. b) Simulation of consolidation in the group with hippocampal lesions. This panel shows the experimental data with error bars, together with a curve describing the buildup of 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.

due to experimenter-determined learning experiences are assumed to be interleaved with ongoing exposure to other patterns, either from the environment or from 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 10a. 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 in the group that received a hippocampal lesion one day after the tone-shock experience. These 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 the background experiences of the animal, which for simplicity we treat as being constant over time. We assume that the neocortical system continues to be exposed to these associations throughout. 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. This 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 exposure to stable aspects of the environment. Although it is one of our assumptions that hippocampal traces decay with time, we ignore this decay for simplicity in this initial simulation. Thus the hippocampal trace remains at full strength for the duration of the experiment (in control animals) or until the hippocampus is removed (for hippocampal groups). This assumption is in accord with the lack of significant forgetting in the control animals.

We monitored the response of the network to each presentation of the new pair, and the performance of the network is graphed in Figure 10b. 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 clearly 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, by taking the mean time spent freezing averaged across the control groups receiving sham lesions at different delays. 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.

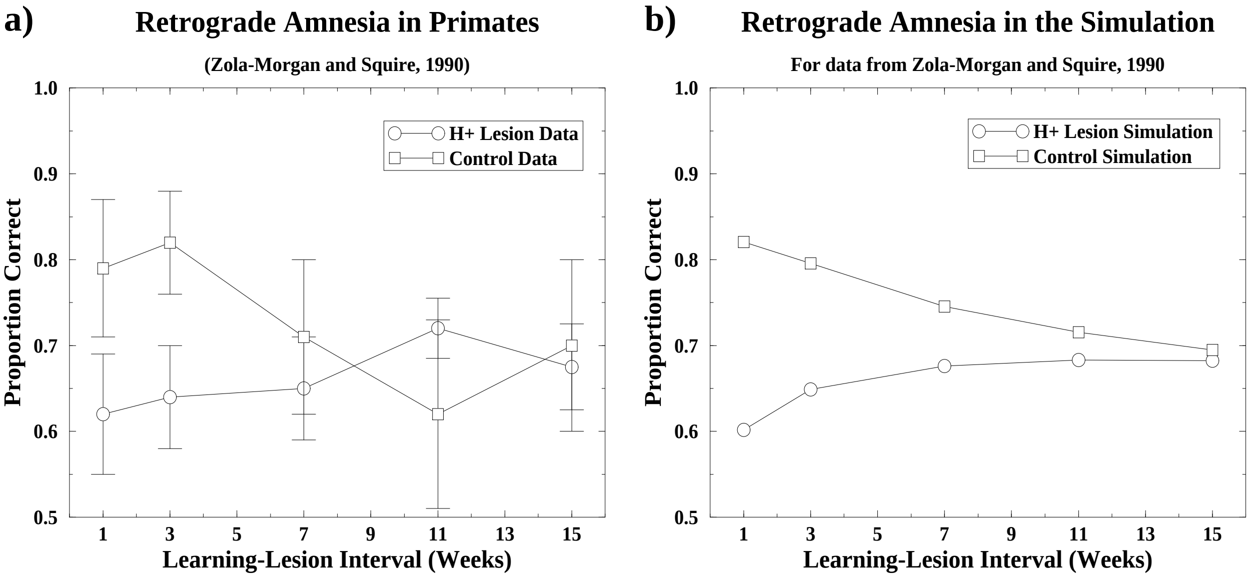

Figure 11: Experimental data (a) and simulation (b) of the experiment of Zola-Morgan and Squire (1990).

*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 11). The hippocampals, 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 hippocampals and the 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; 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 trails 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 cortical 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 $\epsilon$. 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 is shown in Figure 11. 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 necessarily meaningful. Therefore it appears reasonable to conclude that these data are consistent with our approach to consolidation, as illustrated by our simulation results. 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 $\epsilon$ 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. For the controls, the fit is less good, due to the high-variability data point at 11 weeks, which falls well below the data points on either side of it for no explicable reason. 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 results 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

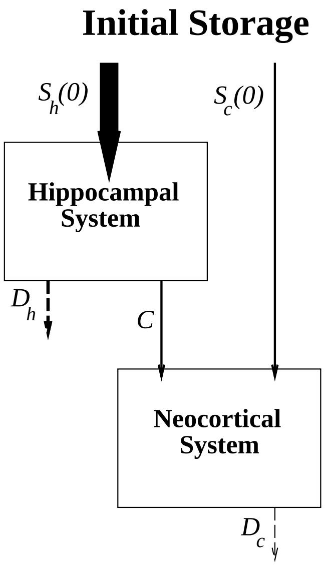

Figure 12: A simple two-compartment model that characterizes memory storage and decay in the hippocampal system, together with consolidation and subsequent decay in the neocortical system. Arrows are labeled with the parameters of the simple model: $S_h(0)$ and $S_c(0)$ refer to the strength of the hippocampal 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.

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. The formulation is depicted graphically in Figure 12.

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:[2]

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

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 will depend on the residual strength of the trace:

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

The parameter $r_h$ designates the reinstatement rate.

[2] 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.

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 $\epsilon$ 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

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

here $C$ is the consolidation rate, equal to the product of $\epsilon$ 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

Table 1: Parameter values used in fitting the simplified two-memory model to consolidation results of four 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.

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 two other studies in Figure 13; 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

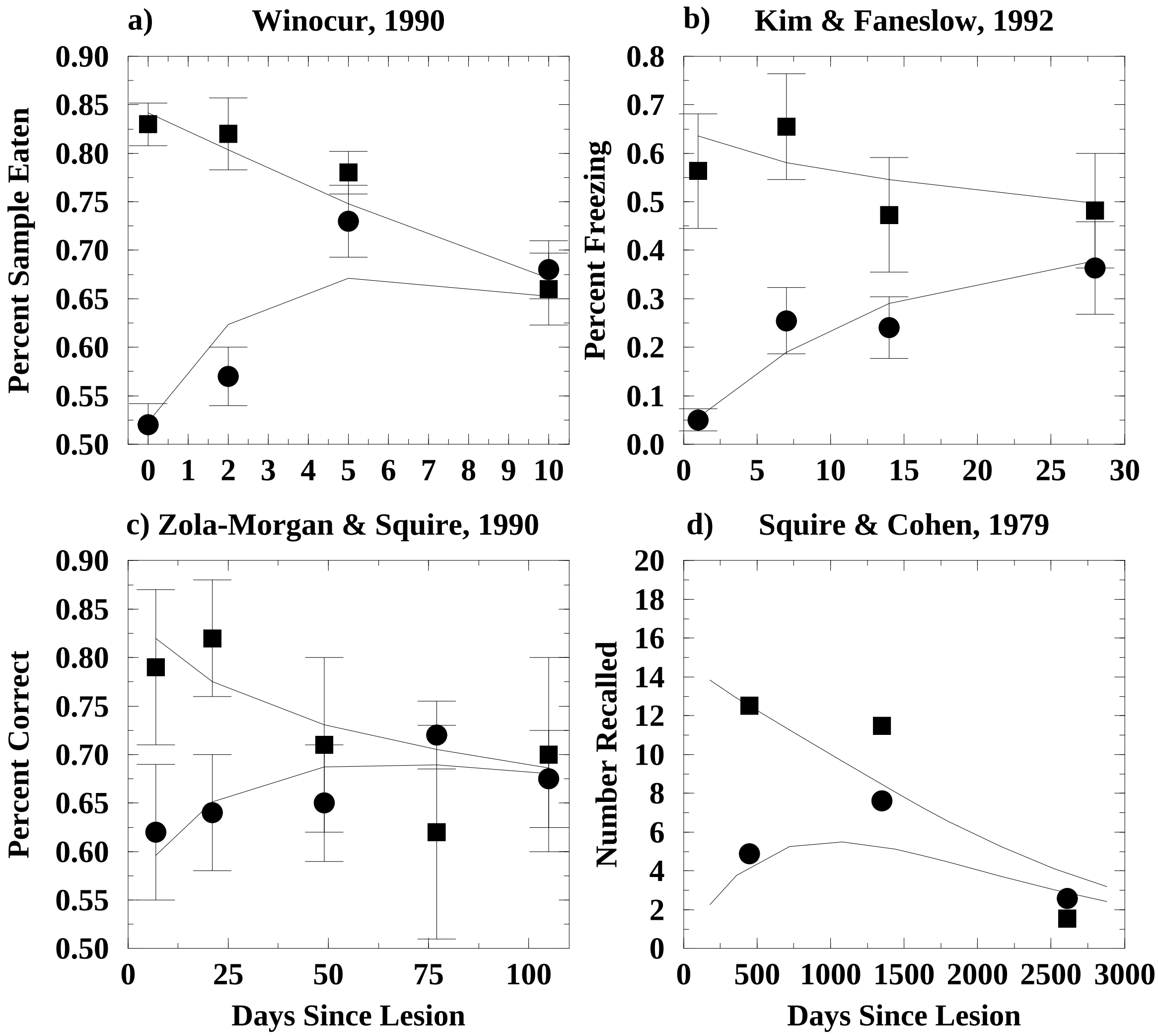

Figure 13: Simulations of retrograde amnesia gradients using the simple two-compartment memory model sketched in the previous figure. Panels (a)–(c) show behavioral responses of animals lesioned different numbers of days after exposure to relevant experiences. (a) Fear behavior shown by rats exposed to a contingency between tone and foot shock in a novel environment (Kim & Fanselow, 1992). (b) Food preferences exhibited by rats exposed to a conspecific (Winocur, 1990). (c) Choices of reinforced objects by rats exposed to 14 training trials with each of 20 object pairs (Zola-Morgan & Squire, 1990). Panel (d) shows recall by depressed human subjects of details of TV shows aired different numbers of years prior to the time of test, either before of after ECT (Squire & Cohen, 1979). Note that we have translated years into days to allow comparison with the results from the animal studies. The parameters of the fitted curves are shown in Table 1.

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 Winocur 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 Electro-convulsive 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.

We chose to concentrate of 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 et al., 1989 and other studies); Squire and Slater (1975) went to some pains 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 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 normal 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 for some reason. 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, though, 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., 1989) 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, indicating a very slow rate of decay of information from the hippocampal formation in human adults.

*Possible Reasons for Variation in the Rate of Neocortical Learning in Different Species*

Overall, the most striking aspect of the retrograde amnesia data presented in Figure 13 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 enough shock to cause marked fear responses. Such a salient experience may result in stronger initial hippocampal traces that show greater resistance to decay.

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. If so, the parameters of the system could be tuned for relatively short persistence in the hippocampus and relatively rapid consolidation.

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 use of extremely small learning rates in humans. However, it should be noted that there are a number of circumstances that can lead to a small *effective* learning rate in connectionist models, even when the actual learning rate is not that small. Thus, differences in apparent learning rates between species could be due to differences in the actual size of connection adjustments, or to the influences of these other factors.

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 in duration (about 3 years) than the retrograde amnesia seen in the older group of patients tested by MacKinnon and Squire (1989) (over 10 years), and suggests that this difference may be due to changes in the rate of consolidation with 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:

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

In this case $e_n$, the estimate of the mean after the $n$th observation $o_n$, is always exactly equal to the mean of the $n$ 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. Interestingly, it is possible that the effective learning rate could decrease with age even if the actual weight-change parameter does not. In particular, networks that have developed strong connection weights after a period of learning can exhibit a lack of sensitivity to new information that does not conform to the structure of the data they have already learned (Munro, 1986). Thus, 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 does not necessarily provide a strong 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, McNaughton, Mizumori, Leonard, and Lin (1990) have found evidence that older animals have poorer hippocampal retention than younger animals. It may then be that 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.

Turning back to the other end of the age dimension, 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. It would also be more difficult to interpret if reinstated, since the reinstated representation would no longer make sense in terms of the more mature system of neocortical representations.

## 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. 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 and describe some ideas that emerge from our account concerning the organization of the hippocampal system and the roles of its various parts.

## *Comparison of our Theory with Other Views of the Role of the Hippocampal System in Learning and Memory*

We are not the first to theorize about 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 deserve attention.

### *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 region 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) 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 teaches 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. The idea of the hippocampus as teacher to the neocortex has also been proposed by Milner (1989). We first discussed the idea in McClelland, McNaughton, O'Reilly, and Nadel (1992) and McClelland, McNaughton, and O'Reilly (1993), and it has recently been adopted by Treves and Rolls (in press). Alvarez and Squire (1994) have also adopted this view. In modeling work done at the same time as but independently of our own, they have developed a neural network simulation of the reinstatement process, showing how it can capture the general form of the pattern of data shown in Figure 1.

While not explicitly concerned with retrograde amnesia, the distinction made by Olton, Becker, and Handelmann (1979) between working and reference memory 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. However, reference memory — memory for that which is constant in an environment — is only one aspect of shared structure. Another example of shared structure of events and experiences may be found in category learning tasks (Knowlton & Squire, 1993; Knowlton, Ramus, & Squire, 1992), and in the kinds of memories that patient HM was able to form for the assassination of President Kennedy or for the layout of his own immediate surroundings (Milner et al., 1968). Amnesic patients can show performance identical to normals in some such tasks (Knowlton & Squire, 1993; Knowlton et al., 1992). We would predict, though, that amnesic patients might well be impaired in the initial acquisition of performance in such tasks to the extent that performance in early stages of practice could be based on memory for particular examples studied. Certainly, patient HM is profoundly impaired in the acquisition of a great deal of information that he must have been exposed to many times; similarly, the rate of acquisition of spatial reference memory knowledge in rats with hippocampal lesions is dramatically impaired, even though eventual performance can reach normal levels (Olton et al., 1979).

Like Olton et al. (1979), many other theorists have focused primarily on the anterograde effects of hippocampal lesions. Many of these theoretical discussions have suggested that the hippocampal system directs the neocortical system in its choice of representations; the neocortical system plays the role of an impoverished learning device that needs the hippocampus to function properly in certain contexts. One popular idea (Rolls, 1990; Schmajuk & DiCarlo, 1992; Gluck & Myers, 1993) has been 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. In these models, the hippocampus plays this role at the time of initial memory formation. Temporally graded and extended retrograde amnesia is therefore problematic for such views: if the hippocampus simply tells the cortex what to learn at the time of the initial encoding event, there is no reason to expect any retrograde amnesia. Wickelgren (1979), whose proposals focused on the role of the hippocampus in constructing new "chunks", did consider temporally graded retrograde amnesia, and offered an account in which the hippocampus was necessary for the initial selection of the neocortical representation and for its subsequent reactivation, until direct intra-cortical connections allow the representation to be activated without the hippocampal contribution. Such an idea could easily be incorporated into the theories of Rolls (1990), Gluck and Myers (1993), and Schmajuk and DiCarlo (1992), and indeed this has happened in the case of Roll's theory (Treves & Rolls, in press). But this move remains *ad hoc* since these theories do not explain why consolidation should necessarily be slow.

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 Squire (1992), Humphreys, Bain, and Pike (1989), O'Keefe and Nadel (1978), Mishkin, Malamut, and Bachevalier (1984), Cohen and Eichenbaum (1993), and Warrington and 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 explicit 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. 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*

Two other aspects of existing theories of hippocampal function deserve further consideration in comparison with our views. At first glance our approach may seem to contrast with these other approaches but on closer inspection the differences may be more matters of emphasis and perspective than substantive points of disagreement.

*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 that is initially apparent (See Alvarez & Squire, 1994 for a related argument). The reason is that our proposal does not require that somehow a full copy of the neocortical pattern of activation is transferred to the hippocampal system. For one thing, reinstatement need not replicate the initial experience in full detail; adequate performance in tasks such as cued recall could occur without a complete reactivation of the full cortical representation that was present at the time of learning. Second, and more important, our view does not require that the hippocampal pattern is a copy of the neocortical pattern. The pattern simply needs to encode enough information about the pattern for the neocortex to reconstruct it. Given the much smaller number of neurons in the hippocampal system relative to the neocortical system, the encoded representation must be highly compressed, but compression need not result in loss of essential information, as long as there is redundancy in the larger neocortical representation. This idea is the basis of compression schemes that are used for computer files, and several neural network models that perform pattern compression have been developed (Ackley et al., 1985; Cottrell, Munro, & Zipser, 1987). In all these cases, the compressed representation can be sufficient to reconstruct the full representation if the latter is constrained and/or redundant. 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 patterns*. A further point of similarity arises from the fact that additional knowledge is needed to implement the pattern compression and decompression processes. This knowledge, we suggest, 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 proper. This reduced representation is then the one that is stored in the hippocampal system. Once this representation is retrieved, 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 detailed information needed to reconstruct a particular pattern of activation is not stored in the hippocampal system.

The remaining issue concerns the acquisition of the information that is not stored in the hippocampal system. One view seems to be that this information is acquired at the time of the experience, and is stored in the neocortex. Squire et al. (1984) raise this possibility. Though we agree that some synaptic modification may take place in the neocortical system at the time of an experience, these changes are only very slight on our account, and thus are not sufficient for storage of much new information in local cortical circuits at the time of initial exposure to a single experience. Instead we suggest that the remaining information is part of the knowledge that has previously been discovered through experience with other inputs. On this view, the encoding and decoding operations are carried out by connections that share the properties of the long-term memory system, and are relatively fixed, at least in adult memory systems. Like the connections within the neocortical system themselves, they would reflect the redundancies and constraints characteristic of the ensembles of events and experiences previously experienced. This proposal makes the prediction that memory will be far superior for materials conforming to the structural constraints that apply to familiar items than it will be for totally random patterns or patterns that embody quite different constraints, since in the latter cases the prior connection weights would not supply the added information and both encoding and decoding would fail. 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).

*Prediction.* A different sort of perspective 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 we believe occurs in the hippocampus. Predication 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. Autoassociative 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 autoassociative completion of material present in experience at overlapping times.

### *Processing within the Hippocampal System*

We have largely treated the hippocampal system as an undifferentiated whole, but in fact it is a highly differentiated system, at least from a structural point of view. The hippocampus proper — consisting of the Dentate Gyrus, CA1, and CA3 areas — has a unique organization, and many authors (e.g., Marr, 1971; McNaughton & Morris, 1987; Treves & Rolls, in press) have proposed unique functions for it, in some cases (Myers & Gluck, 1994; Eichenbaum et al., in press; Murray, Bachevalier, & Mishkin, 1989) differentiating these from the functions of the parahippocampal region, here taken to include the entorhinal cortex, the subicular complex, the perirhinal cortex, and the parahippocampal cortex. Whether some of these regions play important roles in the neocortical system quite apart from their role in learning an memory as parts of the hippocampal system remains a controversial question.

Our own theory of this matter will be presented in detail in a subsequent article. For now, we note briefly two key functional considerations that arise from the fact that the hippocampus plays a role in memory reinstatement and consolidation that extends over a period of years:

- To facilitate use of hippocampal memories over an extended period of time, the system as a whole requires stable and efficient bi-directional communication between the neocortical system and the hippocampal system.

- To allow memories acquired sequentially to persist for long periods, even as new memories are added subsequently to the system, the hippocampal system must use a form of representation that minimizes the interference between the traces of different memories.

We suggest that these two points are addressed by assigning different functions to the hippocampus proper and the parahippocampal region.

*The parahippocampal region.* The primary role of the parahippocampal region, in our view, is to provide an interface between the hippocampus proper and the neocortical system: Stable, bi-directional connections between regions of the neocortical system and the entorhinal cortex (some direct and some mediated by neurons in the parahippocampal and perirhinal cortices) are used to implement the pattern compression function discussed in the previous section. This allows the hippocampus proper to process an appropriately reduced pattern of activation, rather than a full replica of the neocortical representation of the event to be stored in memory. Similarly this allows the hippocampus to generate such a compressed representation in its output, without the need to be concerned with how this compressed representation is translated back into a pattern of activation distributed widely throughout the neocortical system. We emphasize the need for the connections that implement the pattern compression and decompression operations to be stable over time, so that memories stored at one point can still be retrieved and appropriately reinstated as much as several years later.

*The hippocampus proper.* In our view, the hippocampus proper is the place where the plastic changes occur that underlie the initial storage of the contents of an experience. These plastic changes, we believe, are changes to the strengths of the connections among neurons activated in the hippocampus during the initial experience. McNaughton and Nadel (1990) review a wide range of evidence consistent with this view, including the fact that synaptic connections between neurons within the hippocampus show associative long-term potentiation, a form of plasticity that may be the basis of learning within the hippocampal system.

One important factor that determines the durability of memories stored through synaptic modification is the extent to which the patterns of activity used to represent different events overlap. Essentially, if the representations of two events make use of some of the same units, then each will depend, for its reinstatement, on changes to some of the same connections that were modified in storing the other. Some of the changes will work at cross-purposes, and the ability to separately retrieve the unique elements of each trace will degrade the more the patterns overlap. Thus, it is useful for each pattern to activate relatively few units, and for the patterns that represent different memories to have the minimum possible number of active units in common. The hippocampus proper appears to use just such sparse, minimally overlapping representations, as illustrated in Figure 14 from (Barnes et al., 1990). The figure shows the firing rate profiles of two neurons in the hippocampus proper (one in area CA3 and one in area CA1) as well as two neurons from the parahippocampal region (entorhinal cortex and subiculum) in the rat. The profiles were obtained as the animal traverses an eight-arm radial maze in search of food reward at the ends of the arms. The parahippocampal neurons show relatively diffuse firing profiles, indicating that these neurons are somewhat active at many locations in the environment. The neurons from the hippocampus proper, on the other hand, fire only when the animal passes through a small circumscribed region in the space. They fire less often, and have very sharply defined receptive fields, so that the representations of different locations in the space have relatively little overlap.

Our proposals appear to suggest that a lesion restricted to the hippocampus proper would produce memory deficits as severe as those that result from more complete lesions of the hippocampal system. However, this is often not the case. In some tasks, lesions of the hippocampus proper, or at most the hippocampus plus the subiculum (Morris, Schenk, Tweedie, & Jarrard, 1990), produce deficits as great as those produced by larger lesions encompassing much more of the hippocampal system. However, in other cases lesions restricted to the hippocampus proper produce much less severe deficits than lesions that include — or are restricted to — the parahippocampal region (Zola-Morgan, Squire, Amaral, & Suzuki, 1989; Murray et al., 1989; Gaffan & Murray, 1992). The literature paints a complex picture of the circumstances under which selective lesions to the hippocampus proper produce relatively mild or severe deficits (See Squire, 1992; Eichenbaum et al., in press for recent reviews). The picture is clouded by a number of factors, not least of which is the difficulty of producing a selective lesion of the hippocampus proper that is also fairly complete.

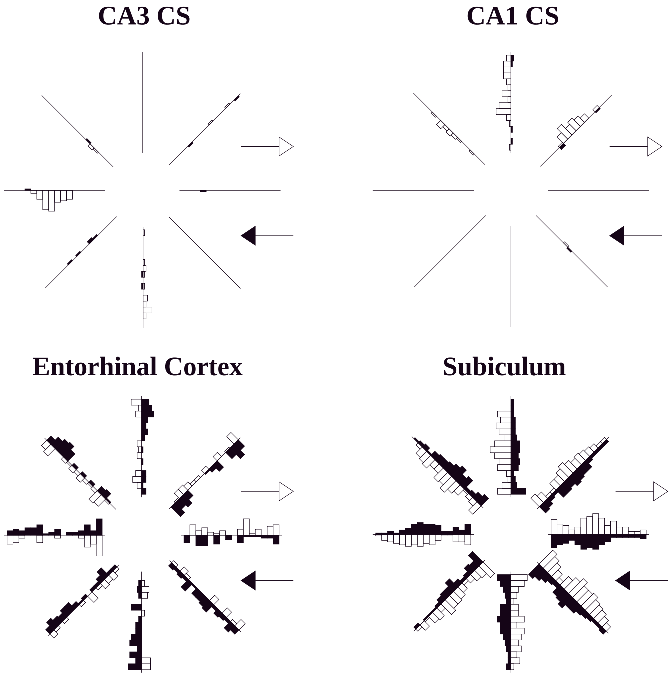

Figure 14: Representative response profiles of neurons in CA3, CA1, entorhinal cortex, and subiculum during performance in a spatial working memory task in the 8-arm radial maze. From Barnes et al. (1990). NOTE: The profiles in the electronic version of this article are redrawn from the originals. The actual originals will be reproduced in the journal publication.

One possibility (Murray et al., 1989; Gaffan & Murray, 1992; Eichenbaum et al., in press) is that the parahippocampal region subserves simple judgments of familiarity based on exposure to a stimulus item in the very recent past (the preceding several minutes). Gaffan and Murray (1992) point out that such effects might be mediated by neurons in this region whose activity is modulated by very recent experience with specific stimuli. This sort of modulation seems unlikely to provide the substrate for the relatively long-term storage of traces of events and experiences that is needed for consolidation, and effects of hippocampal lesions do become apparent in these tasks when longer delay intervals are used. Thus, while there may be some plasticity elsewhere in the hippocampal system relevant to performance in some (short-term) memory tasks, it remains possible that connection changes within the hippocampus proper are the indeed crucial for the initial formation of the memory traces that necessary for consolidation.

## 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 that 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. It may answer the two questions we have posed in this article, but in so doing raises many new ones. Answering these questions will depend on the emerging synthesis of computational, behavioral, and neurophysiological investigation.

# References

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. manuscript.

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.

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

Barto, A. G. & Jordan, M. I. (1987). In *Proceedings IEEE International Conference on Neural Networks*, Vol. 2, 629–636.

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

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). *Journal of Neuroscience*, in press.

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

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.

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. In *Proceedings of the 9th Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Lawrence Erlbaum Associates, 462–473.

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.* Palo Alto: Morgan Kaufmann.

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. (in press). Two functional components of the hippocampal memory system. *Behavioral and Brain Sciences.*

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

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.* San Mateo, CA: Morgan Kaufmann Publishers, Inc., 3–9.

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

Glisky, E. L., Schacter, D. L., & Tulving, E. (1986). 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.

Graff, 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.

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.

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

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

Hinton, G. E. & McClelland, J. L. (1988). Learning representations by recirculation. In D. Z. Anderson (Ed.), *Neural information processing systems.* 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. Cambridge, MA: MIT Press/Bradford, Chap. 3.

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.

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.

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. (1990). The self-organizing map. In *Proceedings of the IEEE*, Vol. 78, 1464–1480.

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

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

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.* Academic Press, 243–305.

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

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

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

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

Marr, D. (1970). A theory for cerebral neocortex. *Proc. Royal Soc 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. (in press). 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.* 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.* New York: Academic Press.

McNaughton, B. L. (1983). Activity-dependent modulation of hippocampal efficacy: Some implications for memory processes. In W. Siefert (Ed.), *Neurobiology of the hippocampus.* New York: Academic Press, 233–251.

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.* Hillsdale, NJ: Erlbaum, 1–63.

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

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.* MIT Press/Bradford, 255–350.

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

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.* San Mateo, CA: Morgan Kaufmann Publishers, Inc., 556–563.

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.* New York: Guilford Press, 65–77.

Morris, R. G. M., Schenk, F., Tweedie, F., & Jarrard, L. (1990). Ibotinate lesions of the hippocampus and/or subiculum: Dissociating components of allocentric spatial learning. *European Journal of Neuroscience*, *2*, 1016–1028.

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. Cambridge, MA: MIT Press/Bradford, Chap. 24.

Murray, E. A., Bachevalier, J., & Mishkin, M. (1989). Effects of rhinal cortical lesions on visual recognition memory in rhesus monkeys. *Society for Neuroscience Abstracts*, *15*, 342.

Myers, C. E. & Gluck, M. A. (1994). Context, conditioning and hippocampal representation. manuscript.

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 Brain Science*, *2*, 313–365.

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. (1993). Generalization with componential attractors: Word and nonword reading in an attractor network. In *Proceedings of the 15th Annual Conference of the Cognitive Science Society.* Hillsdale, NJ: Erlbaum, 824–829.

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

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

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.* San Diego, CA: Academic Press.

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.* San Diego, CA: Academic Press, Chap. 21.

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. Cambridge, MA: MIT Press, 318–362.

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.* Cambridge, MA: MIT Press, 3–30.

Scalettar, R. & Zee, A. (1986). A feed-forward memory with decay (Technical Report NSF-ITP-86-118). Department of Physics, University of California, Santa Barbara, CA 93106.

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

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

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

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.*

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.* Hillsdale, NJ: Erlbaum, 185–210.

Squire, L. R., Haist, F., & Shimamura, A. P. (1989). 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. & 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.

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.

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

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

Treves, A. & Rolls, E. T. (in press). A computational analysis of the role of the hippocampus in memory. *Hippocampus.*

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.

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

Winocur, G. (1990). Anterograde and retrograde amnesia in rats with dorsal hippocampal or dorsomedial thalmic 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, L. M. & 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., & Suzuki, W. A. (1989). Lesions of perirhinal and parahippocampal cortex that spare the amygdala and hippocampal formation produce severe memory impairment. *Journal of Neuroscience*, *9*, 4355–4370.