ANNALS OF THE NEW YORK ACADEMY OF SCIENCES

Volume 843

NEUROSCIENCE OF THE MIND ON THE CENTENNIAL OF FREUD'S *PROJECT FOR A SCIENTIFIC PSYCHOLOGY*

Edited by Robert M. Bilder and F. Frank LeFever

The New York Academy of Sciences
New York, New York
1998

Complementary Learning Systems in the Brain

A Connectionist Approach to Explicit and Implicit Cognition and Memory

JAMES L. McCLELLAND[a]

Carnegie Mellon University and the Center for the Neural Basis of Cognition, Pittsburgh, Pennsylvania, USA

The scientific perspective I will be considering in this article shares several things with Freud's *Project for a Scientific* Psychology. One of these is the effort to build a strong bridge between physiology and cognition. This is clearly a major theme in Freud's scientific program, and it relates very closely to the current emphasis in cognitive neuroscience on building connections between the cognitive and the physiological levels. This article will illustrate one contemporary approach to making such connections. Another thing I myself share with Freud is a primary commitment to the psychological issues. The physiology is very important because it offers a chance to become explicit about the mechanistic basis of psychological processes, but the psychological level—experience, behavior, and cognition—remains the main focus. The big advantage we have today is the opportunity to implement our ideas as explicit models and test their ability to actually affect behavior. If Freud had had this opportunity, I believe that he would have embraced it avidly.

A third point of similarity is that Freud's ideas were very important in making explicit a separation between conscious thought processes and other kinds of aftereffects of experience that influence behavior. His goal was to understand how past experience could influence behavior without explicit, conscious reference to that particular experience. The work that I will be describing relates very closely to this theme.

Finally, the work I will present provides a way of thinking about the nature and function of aspects of sleep processes, including dreaming. The view of the role of sleep is very different from Freud's view, and a little closer to other views expressed in this volume, but there is a degree of common ground in the idea that dreaming provides an opportunity to replay aspects of experience.

The plan of this article is to introduce the concept of complementary learning systems in the brain. We will see that one of these learning systems is really closely tied with what we think of as our explicit memory—our specific ability to remember particular episodes in advance and to relate our current experiences to previous experiences that we and others might have had. The other learning system (or set of learning systems) is one that has much more to do with what we now call implicit learning—learning that influences behavior but without that influence being explicitly noted by us at a time that the behavior occurs.

I will present the neuropsychological evidence that suggests there is a very, very strong dissociation between these two kinds of learning and I will also provide one sort of putative account of how brain systems might be organized so as to implement these

[a] Address for correspondence: James L. McClelland, Department of Psychology, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213. Phone, 412/268-3157; fax, 412/268-5060; e-mail, jlm@andrew.cmu.edu

two very different kinds of learning in the brain. Then I will suggest why the system might be organized in this way. The goal here will be to indicate why there might be a need to have a separation between explicit recollective memory and more implicit learning mechanisms. The article will conclude with a few words about the relationships between ideas and some issues that may be of some clinical relevance, such as how experience becomes ingrained in our immediate psychological reactions to events and experiences, and in our behavioral responses to these experiences.

THE AMNESIC SYNDROME

The starting point for our consideration of the idea of complementary learning systems is the striking neuropsychological dissociation between explicit and implicit learning that is found in patients like HM (Scoville & Milner, 1957). As is widely known, HM underwent a bilateral removal of the hippocampus and related areas on both sides of the brain, after which he totally lost the ability to form new explicit memories. He was unable to remember events and experiences that occurred after the surgery or to explicitly remember having seen a particular individual before or learn the name of someone whom he had met subsequent to the surgery. Within psychological experimental research, the deficit manifests itself quite clearly in tasks such as paired associate learning. Normal subjects can learn to associate randomly paired words with each other, such as LOCOMOTIVE-DISHTOWEL, if they are given several exposures to this pair along with several others in a list. After several exposures, subjects can typically report the second member of each pair, then give the first word of the pair as a cue. In the case of HM, many repetitions of even a short list of paired associates left him virtually unable to recall any of the associated items. Thus, his is a very striking and dramatic deficit.

At the same time, the ability to acquire new cognitive skills, and to show implicit aftereffects of specific experiences remains completely intact. One of the first demonstrations of this was HM's ability to learn how to trace a figure while looking at his own hand doing the tracing through a mirror. These sorts of skills are learned slowly by normals and by amnesics; there is apparently no difference in the rate of learning between normal and amnesic groups (Cohen & Squire, 1980). One of the most striking things about these tasks is that when one brings the patient back, day after day, he or she, despite getting better and better just like normals, cannot remember doing the task, seeing the apparatus before, or knowing what we are going to do with it. Amnesic patients are completely unaware of any prior experience and cannot answer these questions, but they immediately show the effects of prior experience as soon as they begin to perform. Similarly striking is a finding called *spared repetition priming.* In this case, the subject is first shown a list of items to read aloud, e.g., words such as WINDOW (Graf, Squire, & Mandler, 1984). Later, at test, the subject is given a fragment of the word (e.g., WIN__) and asked to complete it with the first word that comes to mind. In this task normals and amnesics complete the fragment with the word previously seen, even though, once again, the amnesics have no recollection of previously reading any words. So, both with skills and item-specific aftereffects, we see a very striking dissociation of implicit learning from explicit memory.

Finally, I want to touch on a third aspect of the neuropsychology that will be very important in our theory of complementary learning systems, namely the phenomenon of a temporally graded retrograde amnesia. This phenomenon was known to Ribot (1882) in the nineteenth century, and has now been replicated in several studies in animals. Basically, the phenomenon is this: If a human or animal has an experience on a given day, and then the hippocampus is removed bilaterally immediately thereafter,

there is a nearly total loss of the memory for that experience. But if the hippocampus is left intact for a period of time after the initial experience and then is removed, the subjects will show gradual increases in the degree of retention. Several studies illustrating this effect are shown in FIGURE 1; one study involves placing the animals in an environment they have never seen before, and subjecting them to pairings of tones with shocks. Experimental animals had the hippocampus lesioned bilaterally, one day,

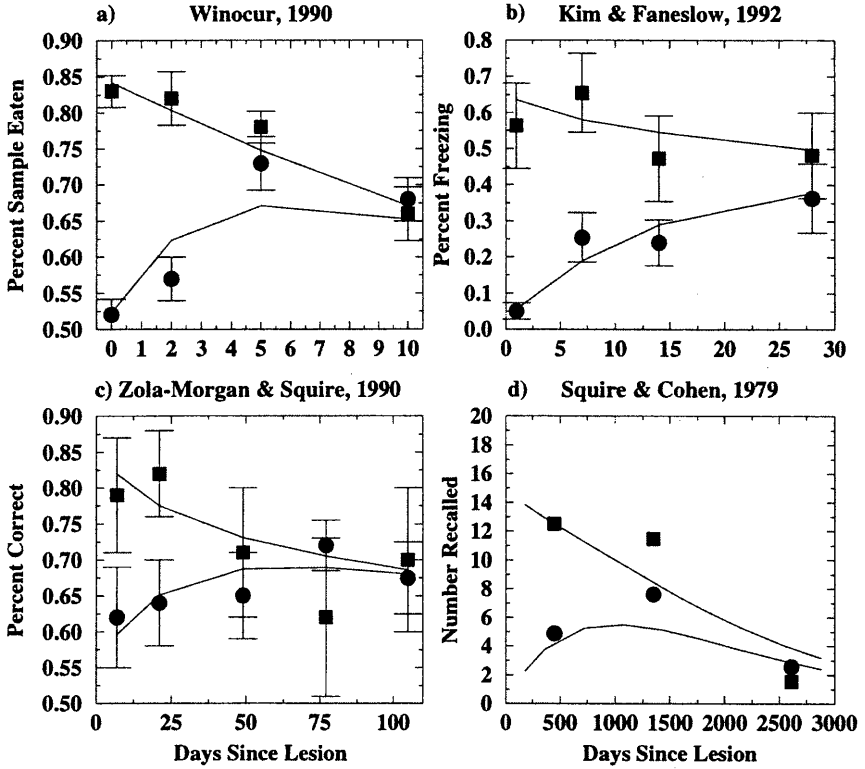

FIGURE 1. Panels **(a–c)** show behavioral responses of animals receiving extensive hippocampal system lesions *(solid circles)* or control lesions *(solid squares)* as a function of the numbers of days elapsing between exposure to the relevant experiences and the occurrence of the lesion. Bars surrounding each data point indicate the standard error. **(a)** Percent choice of a specific sample food (out of two alternatives) by rats exposed to a conspecific who had eaten the sample food. **(b)** Fear (freezing) behavior shown by rats when returned to an environment in which they had experienced paired presentations of tones with foot shock. **(c)** Choices of reinforced objects by monkeys exposed to 14 training trials with each of 20 object pairs. **(d)** Recall by depressed human subjects of details of television shows aired different numbers of years prior to the time of test, after electroconvulsive treatment *(circles)* or just prior to treatment *(squares)*. Here we have translated years into days to allow comparison with the results from the animal studies. NOTE: Data in **(a)** are from Figure 2 in Winocur (1990). Data in **(b)** are from Figure 2 in Kim & Fanselow (1992, p. 676). Data in **(c)** are from Figure 2 in Zola-Morgan and L. R. Squire (1990, p. 289). Data in **(d)** are from Figure 1 in Squire & N. Cohen in (1979). Adapted with permission from Figure 1 in McClelland, McNaughton & O'Reilly (1995). Copyright 1995 by the American Psychological Association.

one week, two weeks, or four weeks after the experience. Control animals had sham surgery at the same time points after the experience. After recovery from surgery, animals were placed back in the apparatus and the extent to which they exhibited fear was assessed by how much time they spent freezing. The animals who had the lesion one day after the experience showed almost no evidence of freezing, whereas those at later periods showed a greater and greater evidence of retention through the freezing measure. The control groups showed strong retention throughout. So, there is some sort of process that goes on in the brain that takes experiences that are initially quite susceptible to lesions of the hippocampal system and then makes them insusceptible. That process has been called *consolidation,* but that is really just a label.

A MECHANISTIC ACCOUNT

The next step in our analysis is to provide a mechanistic characterization of the processes that give rise to the pattern of findings just discussed. The account is not the only one that is possible. It is a theory in the sense that it is an interpretation of the facts. There are a lot of aspects of it that have to be considered hypothetical, but it appears to be consistent with what we know physiologically. There are many who would argue for other interpretations, but this one provides one way of making sense of the data, and will serve as the basis of the rest of what I will have to say. The idea is best described in relation to the schematic diagram shown in FIGURE 2.

The starting point of the account is the idea that information processing takes place via the propagation of activation among neurons in what I called the neocortical system, which includes relatively peripheral input and output systems, as well as more central, highly interconnected brain regions, as illustrated in the figure. Whenever a stimulus is presented to our sensory system, say a visually presented word, it produces a pattern of activity in the appropriate input pathway, in this case the visual one. This activity gives rise to activity in the more central parts of the cortical system, including those perhaps representing the visual appearance, the meaning, and the sound of the word; and this in turn may give rise to an overt response, such as reading the word aloud. In general, any given event or experience produces a rather distributed pattern of activity in many parts of the cognitive system, and the information processing that

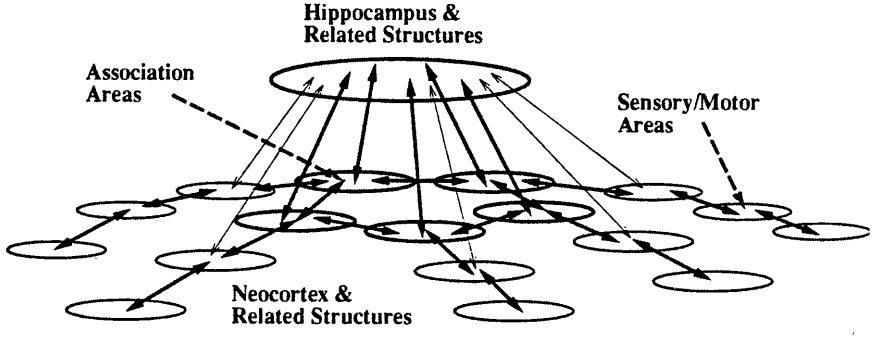

FIGURE 2. A schematic diagram of the information processing and learning systems in the mammalian brain and their interconnections with the hippocampus and related structures in the medial temporal lobes.

we do when confronted with this stimulus occurs through the propagation of this activation.

Now, in this context we think about implicit learning as being the result of small changes that occur to the synapses among the neurons that participate in the processing event itself. In connectionist models, as I would suggest in the brain, these small changes tend to facilitate the processing of the item if it is presented again at a later time. Correspondingly, in the case of mirror reading where the subject is reading one word after another reflected in a mirror, each of those acts of reading a word would produce small changes in the connections that would accumulate gradually, leading to the build-up of the skill of reading mirror-reflected text. A key aspect of the account of the neuropsychological data is the suggestion that the changes that are made on any given processing episode or event are very subtle, and as such they are insufficient to serve as the basis forming adequate associative links between arbitrarily paired items that have never been paired together before.

The pairing of words like LOCOMOTIVE and DISHTOWEL or a particular individual's name with the corresponding face is such a completely arbitrary one, and the notion is that although the kinds of changes that underlie skill-learning are formed with such items as well, those are not sufficient in and of themselves to provide a basis for the rapid formation of an arbitrary association linking these previously unrelated elements. My account proposes that these links rely on the hippocampus and the related structures in the medial temporal lobes of the brain. The notion is that the pattern of activity that is present over the neocortex at the time when a particular experience occurs, such the joint presentation of LOCOMOTIVE and DISHTOWEL by a particular person at a particular time and place, gives rise to a corresponding representation, over a smaller number of neurons, within the hippocampal region. Large plastic changes among the activated neurons within the hippocampal region then serve to store the associations between the particular unrelated elements conjoined within the event. The changes are assumed to occur within the hippocampal system itself. Later on when the word LOCOMOTIVE is presented and the subject is asked to recall the word that had been paired with it earlier in training, the instructions lead to the reinstatement in the cortex of a representation of the training context together with the word LOCOMOTIVE. These in turn lead to a reinstatement of the corresponding aspects of the pattern for the entire event within the hippocampal system. The account assumes that the plastic changes among the neurons there that occurred during learning will allow the rest of the pattern that was present at the time of study to be filled in. This is the crucial step in the recall of the paired-associate DISHTOWEL. The return connections are assumed to play back the results to the cortical system. Recall will not in general be a full reinstatement of the exact details of the original event, but the assumption is that it can be sufficient to provide the basis for an explicit report of the associate.

It should be obvious how these ideas predict that the removal of the hippocampus and related structures would lead to the inability to form new arbitrary associations or to recall arbitrary associations formed shortly before removal of the hippocampal system. At the same time it should be clear why this would not disrupt the subtle item-specific aftereffects seen in repetition priming studies or the gradual development of skills. What remains is to account for temporally graded retrograde amnesia. The suggestion is that there must be some sort of off-line process that reinstates patterns of activity that arose within the hippocampal system during the initial experience, and plays them back to the cortical system. In the case of the Kim and Fanselow (1992) experiment with rats who experienced tones paired with shock in a novel environment, the suggestion is that while the animal is sitting quietly in its home cage or perhaps during sleep, patterns representing the association of tone and shock with the novel environ-

ment are reactivated in the hippocampus and played back to the cortex, and each time the pattern is played back it serves as a kind of training trial for the cortical system. Gradually over repeated reinstatements the cortex learns the associative relationship in the same way it would learn a cognitive skill: by this gradual build-up of small connection changes.

EVIDENCE FROM NEUROPHYSIOLOGY AND ANATOMY

As previously noted, the account I have just given is clearly provisional; it is not the only one consistent with the behavioral data, and there are various other proposals differing in some ways from what I have just described. In what follows, however, I will adopt it as the basis for raising further points of discussion. Before doing this, it is perhaps worthwhile to call attention to some of the exciting physiological data that support several key aspects of the account. For a more detailed presentation of this evidence, see McClelland, McNaughton, and O'Reilly (1995).

Long-term Potentiation

First, it is worth noting that the phenomenon of long-term potentiation was first described (Bliss & Gardner-Medwin, 1973; Bliss & Lømo, 1973) in the hippocampus, and its associative character was established through further research on plasticity in hippocampal pathways (McNaughton, Douglas, & Goddard, 1978; Levy & Steward, 1979; Barrionuevo & Brown, 1983). This form of synaptic modification may be the neural basis of hippocampal-system–dependent memory trace formation. Long-term potentiation (LTP) is extremely easy to induce in several parts of the hippocampus and there are huge number of NMDA receptors there, which are thought to be critical for synaptic plasticity, suggesting as the theory requires, that is a major locus of associative learning in the brain.

Input and Output Pathways

The necessary input and output pathways to carry information into and out of the hippocampal system clearly exist. Squire, Shimamura, and Amaral (1989) document the fact that there are extensive bidirectional projections between nearly all of the neocortical association areas and many other forebrain areas on the one hand and the regions surrounding the hippocampus itself on the other. These para-hippocampal regions (including the entorhinal cortex) in turn project to the hippocampus proper.

Reinstatement of Neural Activity during Sleep

A growing body of data from studies in rats show that patterns of neural activity that arise as the rat locomotes through a spatial environment are re-instated during subsequent sleep (Wilson & McNaughton, 1994). Much more research needs to be done, but the most recent data (Skaggs & McNaughton, 1996) clearly suggest that the location-dependent, sequential patterns of firing seen in the hippocampus during behavior recur while the animals are sleeping after the behavioral session, but not when they are sleeping before the behavioral session starts.

QUESTIONS ARISING FROM THE MECHANISTIC ACCOUNT

Given all of this evidence consistent with our provisional account, we can ask the question, why might the brain be organized in this way? In particular, two questions arise:

- Why do we need this special system in the hippocampus at all, if ultimately all kinds of things are going to be consolidated into the cortex anyway? Why shouldn't the brain have been designed so as to just go ahead and store everything directly in the cortex in the first place?
- Why does consolidation take such a long time? In the rodent, the process can take as much as a month; in humans the evidence suggests that the consolidation process can go on over a period of many years, at least a decade or more. The effects of hippocampal lesions can produce retrograde amnesia gradients that extend over a 10- or 15-year period in humans. So, it looks as though there is a very, very gradual process of consolidating new information initially stored in the hippocampal system into the neocortex.

In the rest of this article I would like to summarize some of the results arising from connectionist models of learning and memory that suggest one set of possible answer to these questions. Connectionist systems are artificial neural networks that abstract away from many of the details of the underlying physiology, but they are neural networks in the sense that they represent currently active mental contents as patterns of activity over a set of simple processing units, and incorporate knowledge and memory in the connections that exist between the units. Connectionist models then allow us to simulate the process of learning in such systems and to explore what happens as we try to teach various things to such systems.

INSIGHTS FROM CONNECTIONIST MODELS

One of the key observations that have come out of connectionist models over the last several years is the fact that they can learn to extract the general structure that is present in ensembles of events in experiences, thereby allowing them to generalize to novel inputs. Chomsky was one of the first to point out that such generalization is crucial. He stressed its importance for language, while others have stressed the importance of generalization for many other domains. The key point for us is that in order for a connectionist model to learn how to generalize properly, it must learn slowly, according to a procedure that I call *interleaved learning.* Basically, the idea is that the system has to learn, not so much the individual cases that make up the examples of a domain, but the general structure of the domain that is exemplified by the individual cases. In order for the direction of change of the connection weights to be governed by the whole ensemble, learning about any case has to proceed very slowly, interleaved with learning about other cases, so that the direction of change is governed by the average over the entire ensemble experiences.

The Domain of Living Things

As an example, let us consider a particular domain in which the idea that knowledge must be structured to support generalization has often been discussed, namely the domain of living things. Quillian (1968) considered living things as an example of the im-

portance of structure in conceptual knowledge. He suggested that there would be great advantages if the knowledge of living things was organized into a hierarchy, with a major subdivision into the plants and the animals and further subdivisions into subtypes such as flowers, trees, birds and fish (note that we speak of a hierarchy based on ordinary experience, rather than biological research in this context). Quillian's proposal is illustrated in FIGURE 3.

The motivation for thinking knowledge is organized as shown in the figure is that this organization allows economical storage of particular factual information at particular points in the hierarchy, and this information can then be used in an inference process to figure out what conclusions apply to particular cases. The hierarchical organization is useful because there are facts that are generally true of everything below a particular node in the hierarchy that will not be true of anything at an alternative node at the same level, so that we can store information economically by organizing our knowledge in this kind of a way. For example, animals move but plants do not, so we can store that at the level of animals. Similarly, birds fly and fish do not, while fish swim and birds do not. We can store those things here and if we wanted to find out whether a robin can fly, we could just follow the links in the diagram: Note that each link represents a proposition, either about a class-inclusion relation ("an X isa Y") or about a specific property or capability ("An X has P" or "An X can C" where P is a property

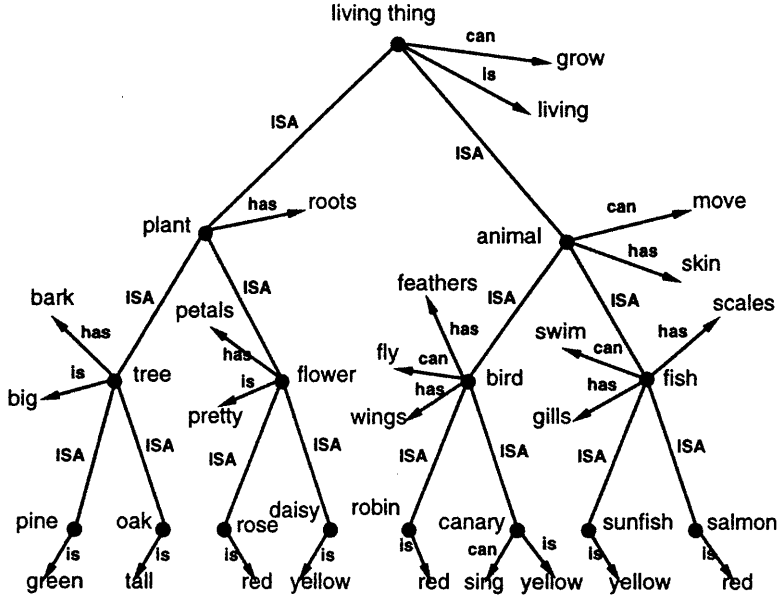

FIGURE 3. A semantic network of the type formerly used in models of the organization of knowledge in memory. All of the propositions used in training the network are based on the information actually encoded in this figure. For example the network indicates that living things can grow; that a tree is a plant; and that a plant is a living thing. Therefore it follows that a tree can grow. All of these propositions are contained in the training set. NOTE: Reprinted with permission from McClelland, McNaughton & O'Reilly (1995). Copyright 1995 by the American Psychological Association.

and C is a capability). A robin is a bird and a bird can fly, so therefore, a robin can fly. Thus, structuring our knowledge provides us with a useful way of storing information efficiently and then using inferences to derive new results. Generalization comes about very simply. If we are told that a new animal is a bird, for example, we can simply store it by connecting it to the bird node via an "isa" link; we can then infer that it can fly just as we could for the case of the robin already considered.

Now, Quillian's particular proposals have fallen from favor for a variety of reasons, and in fact it is not my intention to suggest that connectionist models are ways of capturing Quillian's ideas in exactly the same form. In fact, they capture his ideas in a very different form—one which, I suggest, will eventually be able to overcome most of the difficulties that the original Quillian proposal failed to resolve. Space prevents a detailed presentation of this alternative way of capturing the taxonomic hierarchy, but should be sufficient to provide a general sense of what is involved. The crucial point for us will be to see how in the connectionist system the structuring of the knowledge is captured in the representations that the system learns to assign to various concepts through interleaved learning.

A Connectionist Model that Learns about Living Things

FIGURE 4 shows a connectionist network that can learn the structure that is present in domain of living things as illustrated in FIGURE 3. I should note here that the example, and indeed the choice of this domain to illustrate points related to those I will be making, comes from the work of Rumelhart (1990; Rumelhart & Todd, 1993). The first set of results are based on a replication of simulations reported in these other sources.

The network seen in FIGURE 4 consists of several layers of units. On the left are input units, on the right are output units, and in between are two sets of what are called "hidden units" in connectionist terminology, in that their states are not dictated directly, but are determined by the learning process. The task the network is given is to take inputs in the form of partial propositions consisting of a concept–relation pair, such as, let us say, "robin can," and then produce an output that indicates what a robin can do, by activating the units corresponding to the actual capabilities of a robin. So, in this case the output should be "grow", "move" and "fly"—this correct or desired output is illustrated by darkening the corresponding units in the figure.

Learning in such a network occurs by presenting an input, propagating activation forward through the network to produce an output, comparing the results to the desired output, and then adjusting each connection weight in the network to reduce the discrepancy between the obtained and the desired result. Propagation of activation depends on the connection weights in the network: Each unit simply adds up the inputs it receives from each unit that projects to it, with each input multiplied by the value of the weight on the connection. If the summed input is positive, the unit tends to take on an activation close to 1; if negative, the activation will be close to 0. The exact value of the activation is a smooth, monotonically increasing function of the summed input, bounded below by 0 and above by 1.0.

The network is initialized with random connection weights, so that at first its output is very weak and completely random with respect to the correct response. Gradually, however, as the learning procedure is applied repeatedly for each training case (in this example, the inputs consist of all of the possible concept–relation pairs; in each case the network is trained to activate all of the valid completions of the pair), the network comes not only to be able to produce the correct answer, but also to capture the structure of the domain.

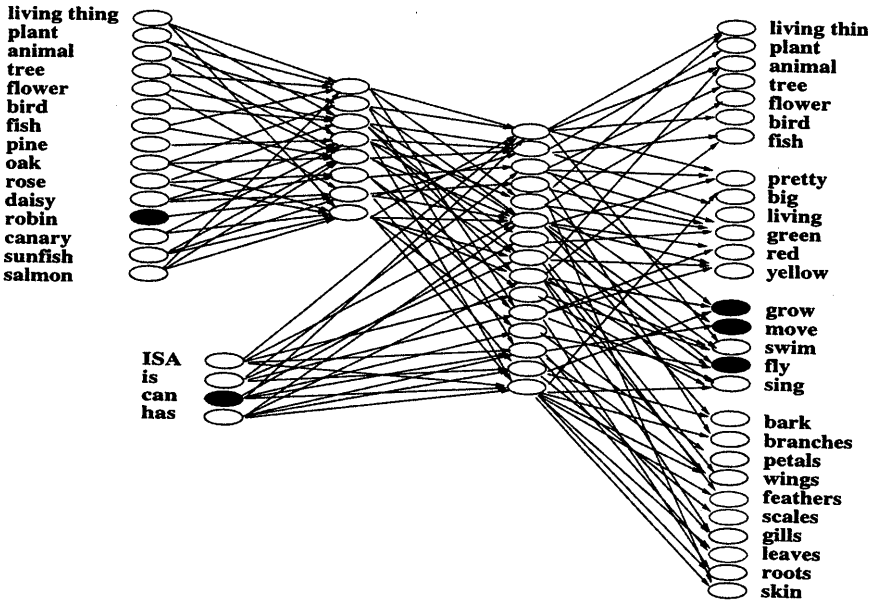

FIGURE 4. Our depiction of the connectionist network used by Rumelhart to learn propositions about the concepts shown in FIGURE 3. The entire set of units used in the actual network is shown. Inputs are presented on the *left,* and activation propagates from *left to right.* Where connections are indicated, 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. NOTE: This figure is based on the network depicted in Rumelhart and Todd (1993, Figure 1.9, p. 15). Reprinted with permission from McClelland, McNaughton & O'Reilly (1995). Copyright 1995 by the American Psychological Association.

Discovery of Structure

The crucial point to note is that it is always possible in such networks to train them to produce any given output in a single trial. If we do that, however, the knowledge will not be structured, and as the network learns each new thing, the changes to the weights will tend to interfere with what it has learned already. On the other hand if the learning is very gradual, so that after each presentation of a particular item the weights are changed just a tiny bit, moving them just slightly in the direction of producing the correct answer, what happens is that the connection weight changes start to build up ways that capture the structure of the domain, so that the network gradually learns to treat as similar those concepts for which the answers are similar. It gradually learns, in short, to assign internal representations inside the network that capture the similarities and differences that exist in the propositions that are true of the concepts.

We can see these points illustrated if we look at the patterns of activation that the network learns to form of the the first (leftmost) set of hidden units, as a result of the

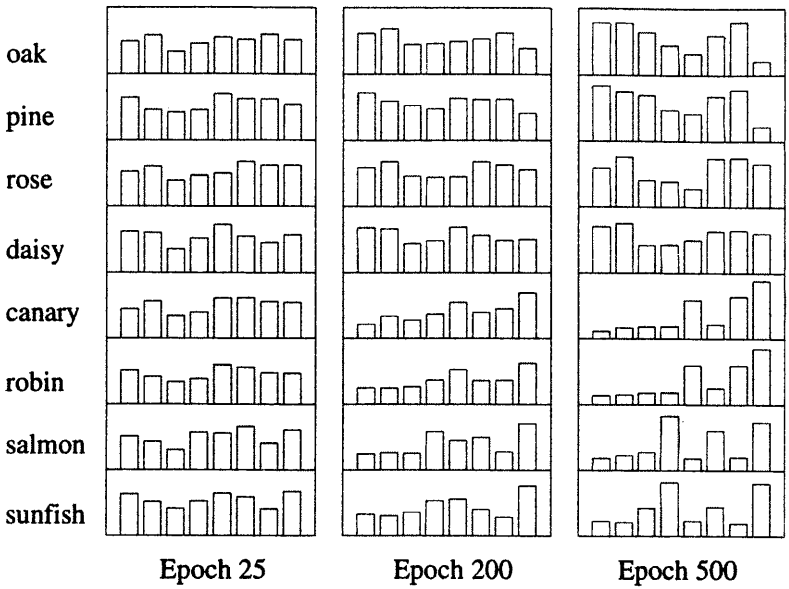

FIGURE 5. Representations discovered in our replication of Rumelhart's learning experiment, using the network shown in FIGURE 4. The figure presents a vertical bar indicating the activation of each of the eight concept representation units produced by activating the input unit for each of the eight specific concepts. The height of each vertical bar indicates the activation of the corresponding unit on a scale from 0 to 1. One can see that initially all the concepts have fairly similar representations. After 200 epochs, there is a clear differentiation of the representations of the plants and animals, but the trees and flowers are still quite similar as are the birds and the fish. After 500 epochs, the further differentiation of the plants into trees and flowers and of the animals into fish and birds is apparent. Reprinted with permission from McClelland, B. L. McNaughton & O'Reilly (1995). Copyright 1995 by the American Psychological Association.

interleaved learning process. These patterns are shown in FIGURE 5 at three different points in learning, one very early in training before the network has learned very much, one at the end of training where it has learned all the facts in this domain, and one at an intermediate point in training. For each concept, at each point in training, eight vertical bars are shown, each of which indicates the activations of one of the hidden units when that concept is presented as the input. What one can see is that early in training all the patterns look very similar. All the units have sort of weak, intermediate states of activation. At the end of training on the other hand, the concepts that are very similar have acquired very similar representations, so the "oak" and the "pine" have very similar patterns of activity as do the "rose" and the "daisy", the "canary" and the "robin" and the "salmon" and the "sunfish." FIGURE 6 shows the same results again, this time in terms of a clustering analysis which essentially recovers, albeit in a different way, the hierarchical structure that Quillian suggested underlies our knowledge of living things. The clustering analysis shows two points that were not as apparent before. First of all, at the end of training, the network has captured the structure of the entire hierarchy, in that the representations of the individual concepts are not only grouped into fish, birds, trees, and flowers, but the fish cluster with the birds and the trees cluster with the flowers. This all occurs because many of the things that are true of fish are

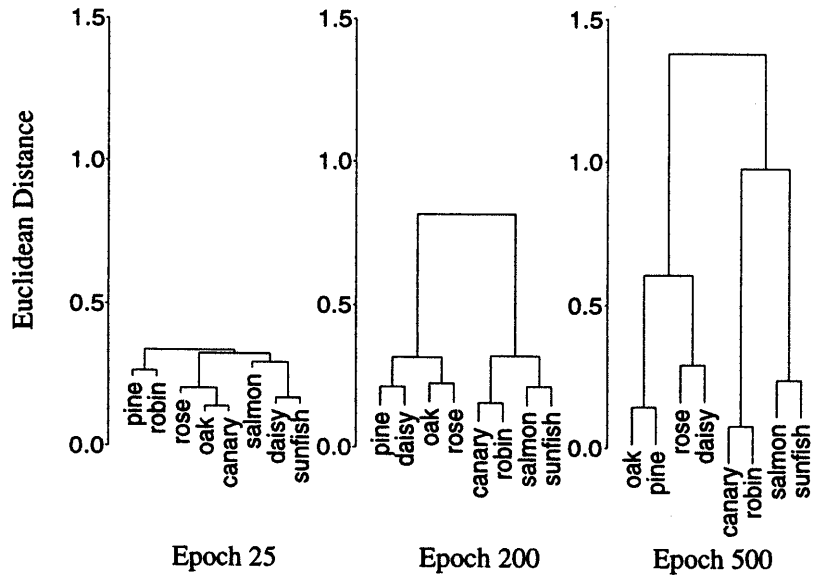

FIGURE 6. Similarity structure discovered in our replication of Rumelhart's learning experiment, using the representations shown in FIGURE 5. These analyses make the similarity relationships among the patterns shown in the preceding figure explicit. The clustering algorithm recursively links a pattern or a previously-linked group of patterns to another pattern or previously formed group. The process begins the pair that is most similar, the elements combined are then replaced by the resulting group, and the process continues until everything has been joined into a single superordinate group. Similarity is measured by the Euclidian distance metric (sum of the squared differences between the activations of the corresponding elements in the two patterns). The height of the point where a subtree branches indicates the Euclidian distance of the elements joined at that branch point. Reprinted with permission from McClelland, McNaughton & O'Reilly (1995). Copyright 1995 by the American Psychological Association.

also true of birds but not flowers and trees, while many of the things that are true of trees are also true of flowers but not birds or fish.

It is worth noting that, while the picture is similar to the one that comes from Quillian's work, the actual representations and processes are quite different. In the present case, the hierarchy is implicit in the similarity relations among the concepts, not explicitly maintained by the use of "isa" links. Although the network can answer "isa" questions, it does not do so by traversing links, but by pattern completion. Category-general properties and capabilities (such as can-fly and has-wings) are not derived by propagating activation up the hierarchy and reading out from a higher node; they are associated with patterns representing each concept. The category-general properties are in general easier to retrieve than concept-specific properties, because they are robustly associated with all of the similar patterns of activation corresponding to all the members of a category, rather than being associated only with those minor aspects of the patterns that differentiate one concept from another. Thus, the connectionist model better captures the fact that category-general properties are usually easier to retrieve than concept-specific ones, as Quillian's model incorrectly predicted.

The second thing that the cluster analysis shows is that the hierarchy emerges from

the top down. Thus, part way through learning, the network has learned to separate the plants from the animals, but has not yet sorted out very well the structure within each category. Within the plants, in fact, the grouping still reflects the random initial weights, rather than the correct, learned similarity relationships. Space prevents detailed discussion, but this property of the network is strikingly similar to the pattern of acquisition seen in development (Keil, 1979). Furthermore, the result of deterioration of semantic knowledge in semantic dementia (Hodges, N., & Patterson, 1995) is consistent with the network's highly robust representation of very coarse-grained distinctions, since such patients quite generally exhibit initial loss of information that differentiates individual concepts and subordinate category membership with a preservation of information that distinguishes concepts in different superordinate categories. Space also prevents illustration of the fact that this sort of learning supports generalization of what is known about some animals of a given type to other similar animals. See Rumelhart (1990; Rumelhart & Todd, 1993) and McClelland *et al.* (1995) for further discussion.

To summarize, then, I have tried to illustrate how the use of a very gradual, interleaved learning strategy allows connectionist networks to learn to represent concepts in a way that captures the structure of the domain in which they are embedded, and that they do so in a way that captures several key aspects of human conceptual knowledge, as well as the development and dissolution of that knowledge.

Catastrophic Interference

While learning structured domains of knowledge is important, humans and other animals do have to be able to learn new information quickly. We can ask, then, what would happen to the structured system of knowledge built up by interleaved learning if we tried to add new information to it, without interleaving. To examine this matter, we considered the case of the penguin, precisely because the penguin is only partially consistent with the knowledge already in the network. While penguins are birds, and have many bird-like properties, they differ from birds in that they swim, and do not fly. In fact I believe that in general novel information tends to be partially consistent with what we already know and to have some arbitrary aspects. The case of the penguin, therefore, exemplifies this general characteristic.

To address this issue, we considered what would happen if knowledge of the penguin were added to the network built up through 500 epochs of gradual, interleaved learning. By means of a focused learning strategy, the network only receives exposure to two training cases: "penguin-isa-bird" and "penguin-can-grow-move-swim." In this case, illustrated in FIGURE 7, the network learns the new information very quickly, but at considerable cost: The training has interfered with what the network already knows about other animals. It now "thinks" that all of the birds can swim but not fly, and that all of the things that can swim but not fly are called birds. This phenomenon, called catastrophic interference, was first illustrated for connectionist models in simulations of paired associate learning (McCloskey & Cohen, 1989) (for related observations, see Ratcliff, 1990). While humans do show some interference in such tasks, it was far less than these networks showed. The results served to call into question the relevance of connectionist models for capturing human learning and memory.

I suggest, on the contrary, that the phenomenon helps to bring out the relevance of these models for our understanding of learning and memory, and in particular for our understanding of the organization of learning systems in the brain. The finding that interleaved learning is crucial for capturing the structure present in a domain of knowledge, and that this sort of learning process appears to capture aspects of conceptual de-

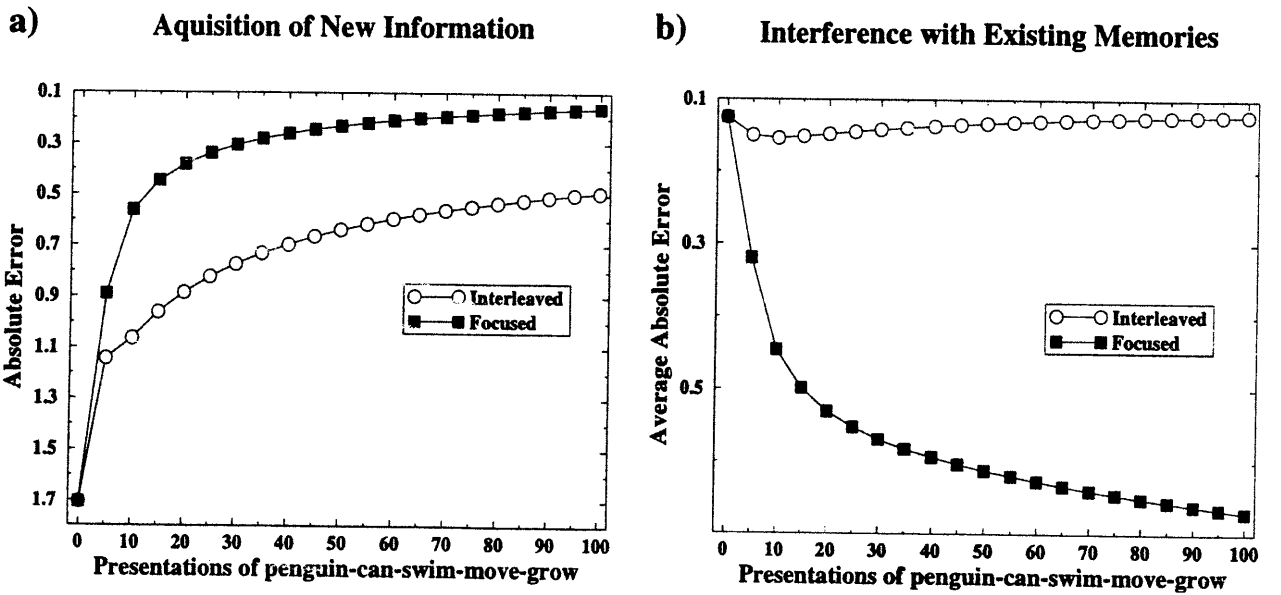

FIGURE 7. Effects of focused and interleaved learning on the acquisition of new knowledge and on interference with existing knowledge. Simulations were carried out using Rumelhart's network, using the connection weights resulting from the initial 500 epochs of training with the base corpus. The performance measure, absolute error, is defined as the sum across output units of the absolute value of the difference between the correct response for each pattern and the actual response. The measure reaches its optimal value of 0 when the output exactly matches the target. Better performance corresponds to lower error, and the axis is inverted for better visual correspondence to standard memory performance curves. In the analysis of interference with other memories, the performance measure is the average of the absolute error over all 15 of the cases in the initial training corpus involving the *can* relation. The scales of each graph are different, and are set to encompass the range of values spanned in each case. The interference is much greater for some items than for others, and falls predominantly on those output units where the correct answer for the pre-existing memory differs from the correct answer for the penguin. Reprinted with permission from McClelland, McNaughton & O'Reilly (1995). Copyright 1995 by the American Psychological Association.

velopment, is consistent with the idea that this gradual, interleaved learning strategy is the one that is used by the neocortex, and with the idea that the cortex makes use of a very gradual learning strategy, allowing only very subtle changes on the basis of a given experience, precisely to support the gradual learning of the structure of entire domains. The finding that attempts to add new information quickly to a system that has gradually learned the structure of a domain through interleaved learning suggests why we need to have a separate learning system, like the one provided by the hippocampus, to allow us to acquire arbitrary new information. Such a system would not only facilitate the initial acquisition of the new information, but, if new memories initially formed in the hippocampus could later be reinstated, this would provide a vehicle that would allow the new information to be gradually integrated into the structured knowledge system in the neocortex.

To illustrate the benefits that would arise from this, we will consider what would happen if the new information about the penguin is interleaved with ongoing exposure to other information about the domain of living things. In this case, the acquisition of the new information takes quite a number of exposures to it, far more than are necessary for focused learning. But this acquisition does eventually occur, and the benefit is that it occurs with only the slightest ripple of interference with the network's existing knowledge of other, more typical birds and fish.

Answers to the Questions

In conclusion, we can come back to the questions raised previously, and see how what has been learned from the study of connectionist models might provide us with answers to the questions.

Why do we need this special system in the hippocampus at all, if ultimately all kinds of things are going to be consolidated into the cortex anyway?

We need a hippocampus to allow the rapid acquisition of new information, without catastrophically disrupting the structured system of knowledge gradually built up through experience.

Why does consolidation take such a long time?

Consolidation is slow to allow the new information to be interleaved with ongoing exposure to other information about the domain so that the new information can be integrated into the neocortical representation. Any attempt at rapid consolidation would produce the very catastrophic interference that the hippocampal system allows us to avoid in the first place.

CONCLUDING REMARKS

The ideas sketched in this article provide both a mechanistic account of the organization of the human memory system and a functional account of the reasons for this organization. A great deal more research is necessary, however, to establish to what degree the arguments presented here are valid. For one thing, considerable controversy remains about the very phenomenon of temporally graded retrograde amnesia. Attempts to replicate the findings have not always been successful, leading to a range of hypotheses about the reasons why there appears to be a temporal gradient in some cases. For another thing, there is considerable ongoing controversy about the mechanistic account itself. Does the learning of new information really take place within the hippocampal region, or, as others have argued, does the hippocampal region serve only to bind together the bits of memories that are actually formed elsewhere? Third, the analy-

sis of the computational models is really in its infancy. The actual learning procedures used in the brain are quite different from those used in the networks discussed here, and though there is reason to think the arguments that have been given here are general to other learning algorithms, it will certainly be necessary to examine in detail just to what extent the arguments carry over to other kinds of networks. Even if the general framework holds up, the detailed workings of the mechanisms that allow memories to be propagated to, stored in, and retrieved from the hippocampus remain to be worked out, as do the details of the circumstances of reinstatement and the resulting neocortical learning. These issues will all need to be resolved before we will have a full understanding of the mechanisms underlying explicit and implicit learning in the brain.

In the meanwhile, I hope this article illustrates two points. The first is the ongoing relevance of some of the themes of Freud's research, especially his insight that cognition is not always governed by explicit knowledge, and his interest in the physiological mechanisms underlying psychological phenomena. The second is that among the technical advances that have increased our ability to address some of these issues is the advent of computational models. These models allow the detailed exploration of the implications of particular mechanistic ideas about the neural basis of cognition in ways that simply were not possible a century ago. It is far from clear where computational models will ultimately lead us, it seems very likely that they will continue to play a role in efforts to make explicit, as Freud tried to do, the mechanistic basis of human behavior, experience, and cognition.

SUMMARY

Freud's ideas about the role of non-conscious processes relates to contemporary thinking about explicit and implicit memory, and his early efforts to understand cognition and behavior in terms of neural mechanisms share several themes in common with contemporary connectionist models. The present paper presents a connectionist perspective of the neural basis of learning and memory and their organization in the brain. The central claim of the article is that the neocortex and many other forebrain learning systems learn slowly so as to become sensitive to the overall structure of experience. Slow learning is crucial for sensitivity to this structure and for organizing specific information with other information in a structured way. The hippocampus and related areas in the medial temporal lobes complement these slow learning systems by providing a mechanism that allows the rapid learning of arbitrary conjunctions of elements that go together to make up an episodic memory.

REFERENCES

BARRIONUEVO, G. & BROWN, T. H. (1983). Associative long-term synaptic potentiation in hippocampal slices. *Proceedings of the National Academy of Science, USA, 80,* 7347–7351.

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

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

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.

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

HODGES, J. R., GRAHAM, N. & PATTERSON, K. (1995). Charting the progression in semantic dementia: Implications in the organisation of semantic memory. *Memory, 3,* 463–495.

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.

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

McCLELLAND, J. L., McNAUGHTON, B. L. & O'REILLY, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102,* 419–457.

McCLOSKEY, M. & COHEN, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.), *The psychology of learning and motivation,* Vol. 24 (pp. 109–165). New York: Academic Press.

McNAUGHTON, B. L., DOUGLAS, R. M. & GODDARD, G. V. (1978). Synaptic enhancement in facia dentata: Cooperativity among coactive afferents. *Brain Research, 157,* 277–293.

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: Appleton-Century-Crofts.

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* (pp. 405–420). San Diego, CA: Academic Press.

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

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

SKAGGS, W. E. & McNAUGHTON, B. L. (1996). Replay of neuronal firing sequences in rat hippocampus during sleep following spatial experience. *Science, 271,* 1870–1873.

SQUIRE, L. R. & COHEN, N. (1979). Memory and Amnesia: Resistance to disruption develops for years after learning. *Behavioral and Neural Biology, 25,* 118.

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

WILSON, M. A. & McNAUGHTON, B. L. (1994). Reactivation of hippocampal ensemble memories during sleep. *Science, 265,* 676–679.

WINOCUR, G. (1990). Anterograde and retrograde amnesia in rats with dorsal hippocampal or dorsomedial thalamic lesions. *Behavioral Brain Research, 38,* 149.

ZOLA-MORGAN, S. & SQUIRE, L. R. (1990). The primate hippocampal formation: Evidence for a time-limited role in memory storage. *Science, 250,* 289