1996.In T.Ono, B.L. McNaughton, S. Molotchnikoff, E.
T. Rolls, and H. Nichijo (Eds.), Perception, Memory,
and Emotion: Frontier in Neuroscience. Oxford:
Elsevier Science, Ltd. 601-613.

39

Role of the Hippocampus in Learning and Memory: A Computational Analysis

J. L. McCLELLAND

Center for the Neural Basis of Cognition and Carnegie Mellon University, Pittsburgh, Pennsylvania, U.S.A.

Human amnesia

A striking finding in the literature on human memory is the profound but selective deficit seen after damage to the hippocampus and related structures in the temporal lobes of the human brain. This pattern of deficits was first seen in patient H.M. (Scoville and Milner, 1957), and has now been replicated in the data of a large number of other patients with similar lesions. These patients show a dramatic deficit in the ability to acquire new explicit memories for the contents of specific episodes and events. For example, the ability to learn arbitrary word-pairs such as "LOCOMOTIVE–DISHTOWEL" is drastically impaired. Normal subjects will learn such lists after a few presentations, so that whenever the first word of a pair is presented they will be able to recall the second word of each pair. But H.M. and other similar amnesics may fail to learn even one pair after repeated presentations. There is also profound impairment in everyday memory abilities, such as the ability to learn names of new people, or the ability to remember important personal events and experiences.

At the same time these amnesics are completely normal in the use of their existing semantic and procedural knowledge, and in fact their acquisition of new skills appears to be completely intact. They also show after-effects from processing particular items that are identical to the after-effects seen in normal people, in a wide range of tasks. For example, if amnesic subjects read a word such as "WINDOW" on a list, they show a normal amount of perceptual facilitation in the identification of visually presented words due to a prior presentation of the word. Amnesics also show a severe temporally graded retrograde amnesia for episodic information. Figure 1 shows data from four

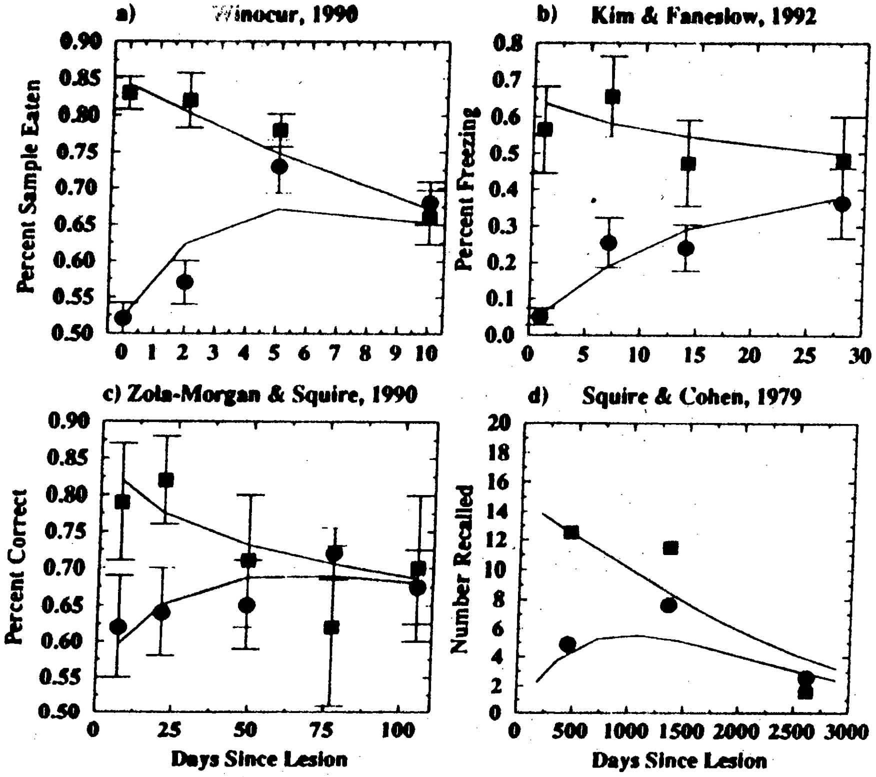

FIG. 1. Panels (a)–(c) show behavioral responses of animals lesioned different numbers of days after exposure to relevant experiences, while panel (d) shows a comparable pattern from an experiment with human subjects. (a) Food preferences exhibited by rats exposed to a conspecific (Winocur, 1990). (b) Fear behavior shown by rats exposed to a contingency between tone and foot shock in a novel environment (Kim and Fanselow, 1992). (c) Choices of reinforced objects by monkeys exposed to 14 training trials with each of 20 object pairs (Zola-Morgan and 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 or after ECT (Squire and Cohen, 1979). Note that we have translated years into days to allow comparison with the results from the animal studies. Reprinted from McClelland *et al.* (1995).

studies of retrograde amnesia in rats, primates, and humans. In all cases we see that there is a selective loss of memory for material occurring shortly before the lesion, relative to more remote time periods. In humans, the period of selective loss can extend over several years, though it is much shorter in rats and monkeys.

Organization of the human memory system

One view of the organization of mammalian memory that is consistent with these facts is illustrated in Fig. 2. According to this view (McClelland *et al.*, 1995), all kinds of knowledge can ultimately be stored in the connections among neurons in the cortex and other nonhippocampal structures such as the basal ganglia (for brevity I will refer to this as the cortical system). We imagine that this cortical system learns gradually, with each experience producing very small changes to connection weights. These small changes, we assume, are the basis of repetition priming effects, and they gradually give rise to cognitive skills. However, in addition to this cortical system, we assume that there is also rapid storage of traces of specific episodes within the hippocampus. On this view, when an event or experience produces a pattern of activation in the neocortex, it produces at the same time a pattern of activation at the input to the hippocampus. So for example, if I am introduced to someone, I form a pattern of activation including the appearance of the person, the name, and other aspects of the situation. Synaptic modifications within the hippocampus itself then autoassociate the parts of the pattern. Later, when a retrieval cue is presented (let us say the person reappears and I wish to recall his name), this then produces a partial reinstatement on the hippocampal input pattern. This is then completed with the aid of the modified synapses in the hippocampus, and then reinstated in the neocortex via return projections. The assumption is that these changes are large enough so that with one or a few trials they may sustain accurate recall, while the changes that take place in the cortex are initially too small for that. Gradually, though, through repeated reinstatement of the same trace, the cortex may receive enough trials with the same association to learn it in the neocortical connections. This reinstatement can occur, we suggest, either through repetition of the association over many different events; or through repeated reinstatement from the hippocampus. Thus on this view the hippocampus can serve both as the initial cite of storage and also as teacher to the neocortex.

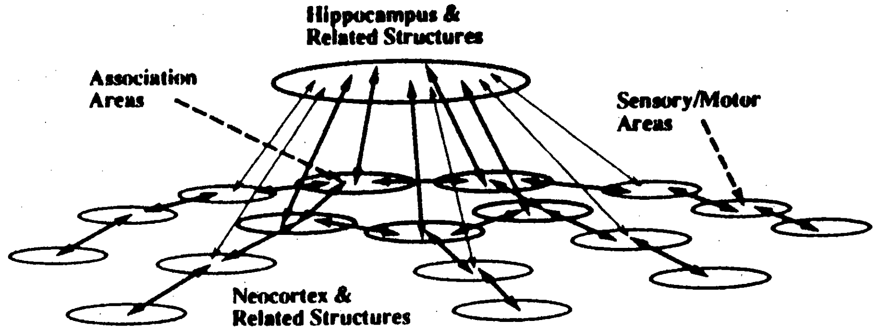

FIG. 2. Conceptual sketch of the neocortical system and its connections with the hippocampal system.

Why is the system organized this way?

The above description provides an account of the data, and although not everyone accepts it, it is not really totally new. Many investigators have suggested some or most of these same points. What I do not know of, however, is any clear explanation of why the system might be organized in this way. We can phrase the questions as follows: First, why is it that we need a special system for rapid storage of memories, if knowledge of all types is eventually stored in connectionist within the neocortex? Second, why is consolidation of information into the neocortex so slow? Why does it require many repeated reinstatements apparently spanning years in some cases?

The answers to these questions require us to consider the acquisition of both procedural and semantic representations in connectionist networks. In such networks, the effective discovery of the shared structure that underlies an entire domain—either a procedural domain such as learning a skill or a semantic domain—requires gradual learning, in which the learning of any one association is interleaved with learning about other associations.

First we consider procedural learning. The network—a fairly standard feedforward network trained with backpropagation—was developed by Plaut and McClelland (1993) to learn to translate alphabetic patterns representing the spellings of words into phonological patterns representing their sounds. The training corpus consisted of a set of 3000 monosyllabic words. Training proceeded very gradually, capturing the gradual nature of the acquisition of reading skill; in fact the network was exposed to the entire corpus of training examples several hundred times, with presentations of more-frequent words given greater weight than presentations of less-frequent words. After training, the network could read 99.7% of the words in the corpus correctly, missing only ten words that were low in frequency and very exceptional either in their spelling or in their pronunciation given the spelling—two examples are the French words "sioux" and "bas", which are very rare and completely inconsistent with English pronunciation. For the vast majority of words, including exceptions such as "HAVE", "GIVE" and "PINT", the network was correct. The network was also tested with many pronounceable nonwords, and scored as well as normals in providing plausible pronunciations—as high as 98% correct on some sets of items. Based on comparisons of the networks performance with typical adult English speakers, we can say that the network has gradually learned both the regular mapping between spelling and sound, and the commonly occurring exceptions to it in a way that corresponds to native English speakers' capabilities in this regard. It has mastered, in short, the procedural skill of translating from spelling to sound—through gradual, incremental learning.

Let us now consider a model that learns semantic information. This model was constructed by Rumelhart (1990) to demonstrate how structured semantic representations can arise in PDP systems from experience with propositions about objects and their properties. The network was trained to capture the information standardly stored in old-fashioned semantic networks of the kind illustrated in the Fig. 3. The goal is to allow the network to store information about concepts in such a way as to permit it to generalize from what it has learned about some concepts to other related concepts. In the old approach, this was done by storing information true of a class or subclass of concepts at the highest possible level of the tree. In the new approach, this is done by gradually learning to assign each concept a distributed representation, capturing its similarity relations to other concepts.

Rumelhart's network is shown in Fig. 4. It consists of a set of input units, one for each concept and one for each of several relations: "ISA", "HAS", "IS", and "CAN". At the output it consists of a set of units for various completions of simple propositions, such as "ROBIN ISA BIRD", "ROBIN CAN FLY", etc. In between there are two layers of hidden units, one to represent the semantic relations of the concepts and another to combine them with the relations to activate the correct outputs.

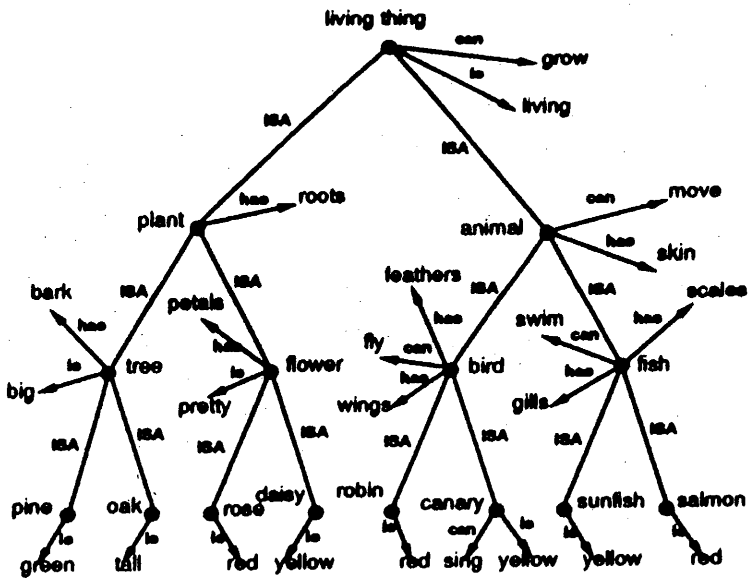

FIG. 3. A semantic network of the type formerly used in models of the organization of knowledge in memory. After Rumelhart and Todd (1993). Reprinted from McClelland *et al.* (1995).

Rumelhart trained this network with a set of propositions involving the concepts shown in the previous figure. There was one input pattern for each concept-relation pair, and the task was to turn on all the output units representing correct completions of the proposition. For example, when the input is "ROBIN ISA" the correct output is "BIRD", "ANIMAL", "LIVING THING". The entire set of patterns was presented to the network many times, making small adjustments to the strengths of the connections after the presentation of each pattern.

After the network has mastered the training set, it is possible to examine the representations that the network has learned to assign to the words in the concepts. We can look either at the patterns themselves, as well as their similarity relations. McClelland *et al.* (1995) repeated this simulation, and the results are shown in Fig. 5. The figure demonstrates that the network has assigned similar patterns to similar concepts—so for example, the pattern for oak is similar to the pattern for pine, the pattern for daisy is similar to the pattern for rose, etc.

Now we may consider how we may achieve generalization in this network. We may do so if we can assign to a new pattern a representation that is similar to the representation of other similar concepts. To illustrate this, Rumelhart trained the network to produce the correct class inclusion output for the concept sparrow. He simply presented "Sparrow isa" as input, trained the network to produce the correct response = "BIRD, ANIMAL, LIVING THING". This caused the network to assign a pattern to sparrow very close to the patterns for robin and canary. As a result the network was able to answer reasonably when probed with other propositions involving sparrows. It said the sparrow can grow and can fly, it has wings, features and skin, and it is small. It was unsure about whether the sparrow could sing and about its color.

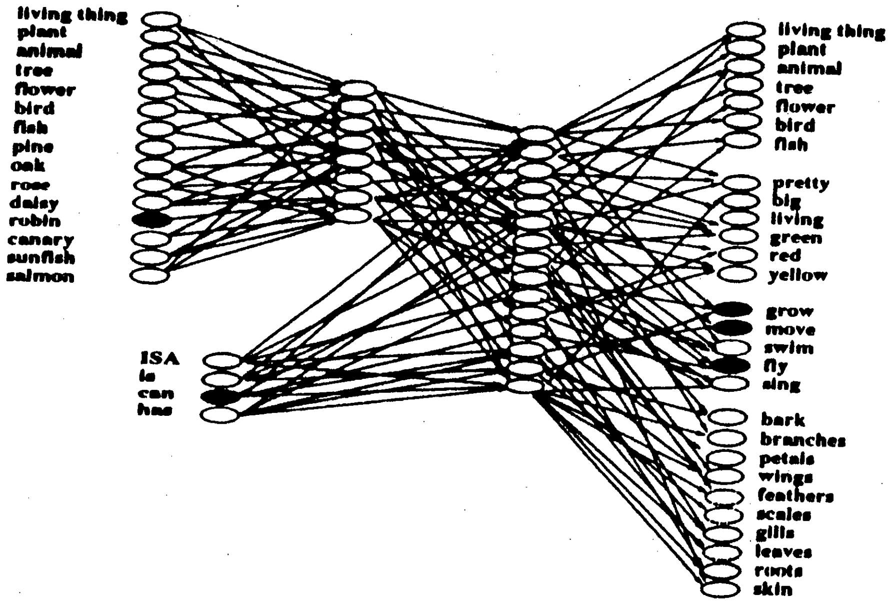

FIG. 4. The connectionist network used by Rumelhart to learn propositions about the concepts shown in the Fig. 3. The entire set of units used in the actual network is shown. Inputs are presented on the left, and activation propagates from left to right. Where connections are indicated, every unit in 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. Reprinted from McClelland *et al.* (1995).

Thus the network learned to inherit knowledge from what it had learned about other birds without direct instruction.

The point here is that PDP models, trained slowly via interleaved presentation on a representative sample of an entire domain of knowledge, can gradually acquire knowledge of structured procedures such as spelling–sound correspondence and structured semantic domains such as the domain of living things.

To make the point even more clear, let us consider what happens in Rumelhart's semantic network if we try to teach it some new information in either of two different ways. The first way, which we call focused learning, involves teaching the network the new information all at once, without interleaving it with ongoing exposure to the structure of the entire domain. The second way, which we call interleaved learning, involves simply introducing the new information into the mix of experiences that characterize the entire domain. As our example, we will consider the case of the penguin. Now as we all know, the penguin is a bird, but it can swim, and it cannot fly. We therefore took Rumelhart's semantic network and taught it these two things in the two different ways just described. The focused case involved simply presenting the two items repeatedly and watching the network learn. The interleaved case involved adding the two new items to the same training set used previously and continuing training as before. The results of these experiments, for the case of "penguin can grow–swim–fly", are shown in Fig. 6. Here if we look at the number of presentations required for acquisition, in the left panel of the figure, we see that focused learning is better, since acquisition is considerably faster with focused learning than it is with interleaved learning. But it turns out that this slight advantage has been purchased at a considerable cost. For it turns out that with focused learning, the training on the case of the penguin has strongly corrupted the models understanding of other concepts. Indeed if we consider the networks knowledge of the concept robin, in the right panel, we see that focused learning has produced a dramatic interference. In fact, the model now thinks that all animals can swim, and even some plants. But in the case of interleaved learning, we see very little interference with what is already known. To be sure there is a slight effect, but as interleaved learning proceeds this is even reduced gradually over trials. Thus with focused learning new knowledge corrupts the structured database in the network; with interleaved learning new knowledge is gradually added with no disruption.

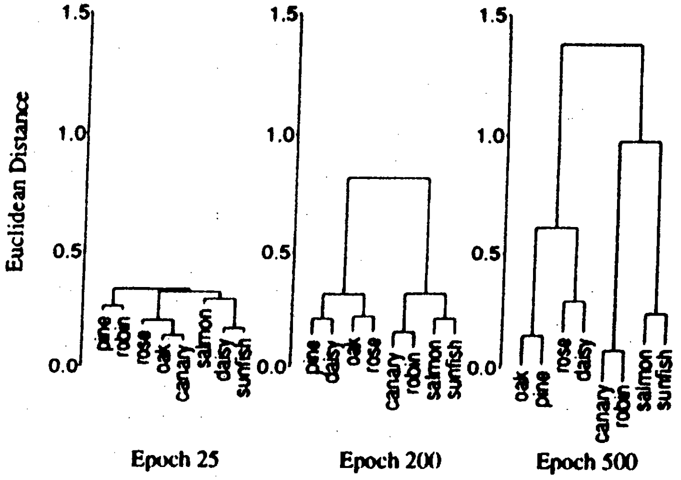

FIG. 5. Similarity structure discovered in our replication of Rumelhart's learning experiment. Initially, the patterns are all quite similar, and the weak similarity structure that exists is random. The concepts become progressively differentiated as learning progresses. Reprinted from McClelland *et al.* (1995).

This demonstration recapitulates a phenomenon that has previously been reported by researchers who have tried to use networks of the type used in Rumelhart's semantic network to model episodic memory (McCloskey and Cohen, 1989). They termed the phenomenon catastrophic interference, and rejected networks of the type that gradually discover structure through incremental learning as models of episodic memory, since storage in episodic memory clearly involves what we have called focused learning. Backpropagation networks are poor models for this kind of learning. However, the kind of learning that they are good at is just the kind of learning we think the cortex is specialized for—gradual discovery of skills from the overall structure of experience.

For the discovery of overall structure, gradual, interleaved learning is necessary, so that the connection weights in the network come to reflect influences of all of the

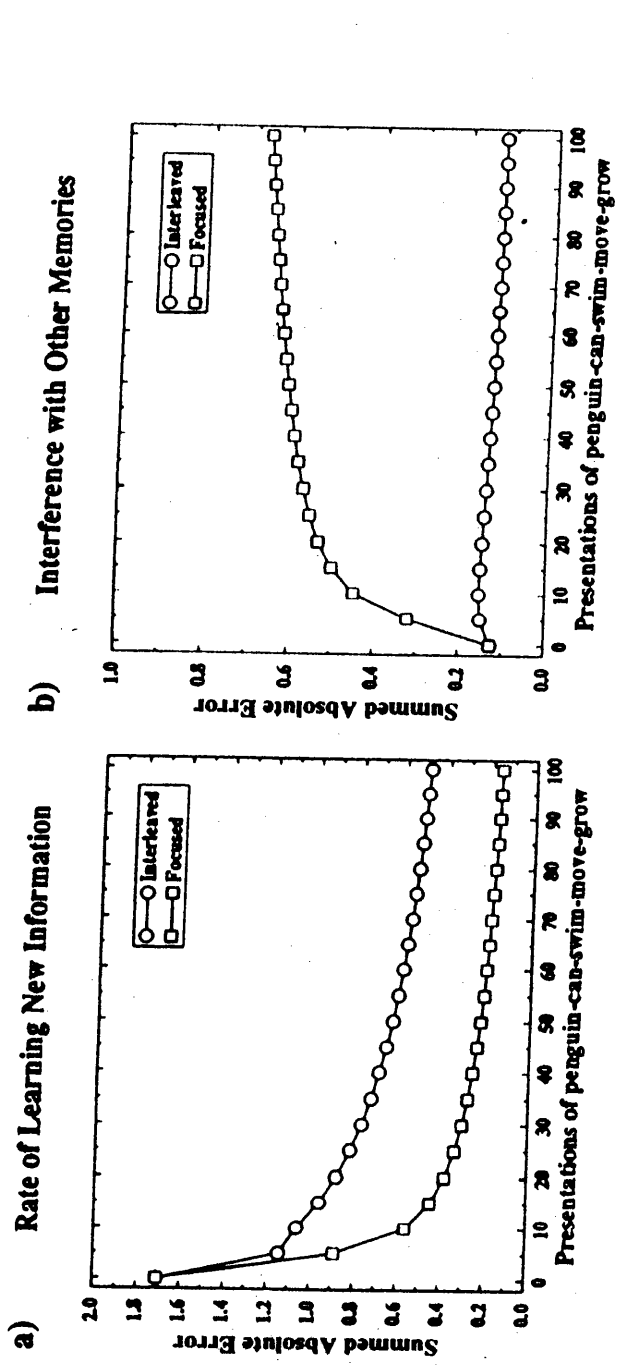

FIG. 6. Effects of focused and interleaved learning on the acquisition of new knowledge and on interference with existing knowledge. Simulations were carried out using Rumelhart's network, using the connection weights resulting from the initial 500 epochs of training with the base corpus. Reprinted from McClelland *et al.* (1995).

examples in the domain. In fact, it can be shown mathematically that in order to converge on a set of connection weights that best captures the structure of a particular domain, it is necessary in the limit to reduce the learning rate asymptotically to 0; the closer it gets to zero, the better we will approximate correct connection weights. This point leads to the observation that if the brain is to be able to extract the structure, it is going to have to learn slowly.

With these ideas in mind, we can now return to the two questions asked above.

First, why is it that we need a special system for rapid learning of the contents of specific episodes and events, if knowledge of all types is eventually stored in connectionist within the neocortex?

The answer begins with the idea that the cortical system is specialized for the extraction of shared structure of events and experiences. In order to do this, and to avoid catastrophic interference with that structure, it is necessary for the cortical system to use a very small learning rate. In this context, the role of the hippocampus is to provide a system in which the contents of specific episodes and events can be stored rapidly, without at the same time interfering with the structure that has been extracted by the cortical system.

Second, why is consolidation of information into the neocortex so slow? Why does it require many repeated reinstatements apparently spanning years in some cases?

Our answer is that consolidation is slow precisely because it would be disruptive if information stored in the hippocampus were incorporated in the neocortex all at once. It is only when new information is gradually interleaved in this way that we are able to incorporate new knowledge into the neocortical system without interfering with what we already know.

Simulation of retrograde amnesia

This observation has now lead to simulations of retrograde amnesia. Due to space limitations I will describe only the simulation of the experiment of Zola-Morgan and Squire (1990). They taught monkeys a series of conditional discriminations. Each conditional discrimination involved two junk objects (for example, a blue plastic cup and a yellow toy dump truck). Under one object the animal could find a reward. The monkeys received 14 trials with each pair of junk objects. Training was arranged so that each monkey learned 20 different discriminations at 16 weeks before surgery, another 20 at 12 weeks before surgery, and 20 each at 8, 4, and 2 weeks before surgery. Some animals received lesions removing the hippocampus and related structures, and others were controls with sham surgery. After time for recovery from surgery the animals were tested and the hippocampal group showed a selective deficit for items learned 2–4 weeks before surgery compared to items learned 12–16 weeks earlier. To simulate this, we imagined that the memory system consists of two parts, a slow-learning cortical network—simulated using a standard backpropagation network—and a more rapidly learning hippocampal network. We do not actually implement a hippocampal network, but instead we treat is at a black box with certain storage, decay, and retrieval properties, and we imagine that it can serve as a source of training patterns for the neocortex. The basic idea (illustrated in Fig. 7) is that when an experience occurs it leaves a trace both in the hippocampus and in the neocortex, but initially the trace in the hippocampus is much stronger. The trace decays relatively rapidly from

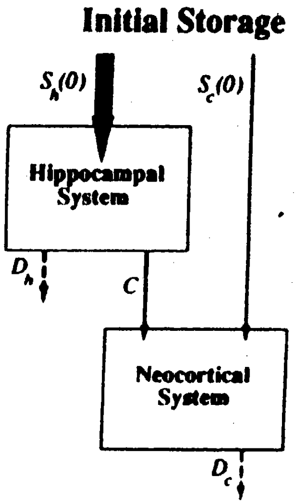

FIG. 7. A simple two-compartment model that characterizes memory storage and decay in the hippocampal system, together with consolidation and subsequent decay in the neocortical system. Arrows are labeled with the parameters of the simple model: $S_h(0)$ and $S_c(0)$ refer to the strength of the hippocampal 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. Reprinted from McClelland *et al.* (1995).

TABLE 1. *Parameter values used in fitting the simplified two-memory model to data from four consolidation experiments*

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

D_h represents the rate of hippocampal decay; C represents rate of consolidation in offline 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.

the hippocampal system, but while it remains it has the chance to be reinstated in the neocortex. Meanwhile in the neocortex other events and experiences are coming along and changing the connection weights as well so the learning of the new experience is interleaved with ongoing exposure to other events and experiences. The simulation results are shown along with the data from Zola-Morgan and Squire (1990) in Fig. 8.

This simulation has led us to consider the parameters of the consolidation process. As noted previously there are vast differences between the rate of consolidation in different studies. Indeed, we were able to fit parameters to the simple consolidation model shown in Fig. 7—the results are shown in Table 1. The fits indicate that the rate

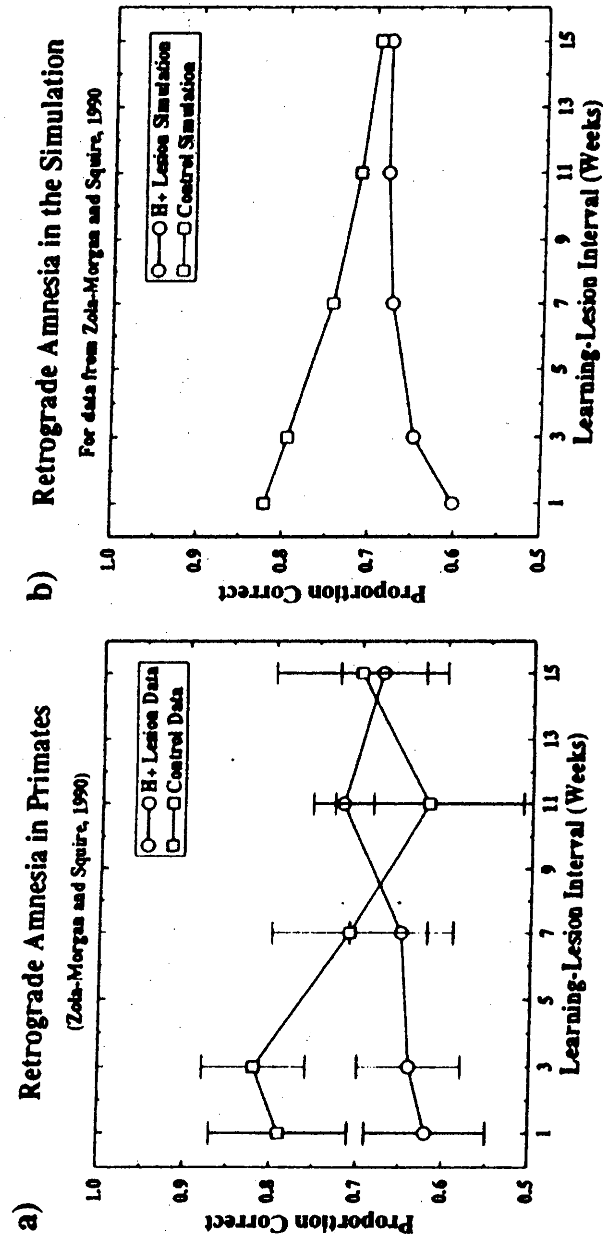

FIG. 8. Experimental data (a) and simulation (b) of the experiment of Zola-Morgan and Squire (1990). Reprinted from McClelland *et al.* (1995).

of decay from the hippocampal system, the rate of consolidation into the neocortical system, and the rate of decay from the neocortical system tend to covary across studies. We consider possible reasons for the differences between species below. Here we consider the difference between the studies of Winocur (1990) and Kim and Fanselow (1992). In the former, the rate of hippocampal decay is apparently far greater than in the latter. The presentation of the shock used by Kim and Fanselow (1992) apparently led to effects that persist for far longer than the experience used by Winocur (1990), who simply exposed rats to a conspecific who had eaten food flavored with cinnamon or chocolate. In the latter case the animals acquired a preference for the food eaten by the conspecific, but even in normals this preference wanes over several days. One possible explanation for this may be that highly emotionally charged experiences such as those involving the kind of intense shock used by Kim and Fanselow (1992) produce much longer lasting hippocampal traces than less emotionally charged experiences like exposure to a conspecific who has eaten a particular food. Interestingly there is evidence (Barnes, 1979; Abraham and Otani, 1991) that strong inducing stimulation produces a longer lasting form of long-term potentiation in hippocampal slices than is produced by weaker stimulation. The decay constants of the weak and strong forms of potentiation in the rat hippocampus are remarkably similar to the decay constants found in our fits to the data from Winocur (1990) and Kim and Fanselow (1992), respectively.

Implications

The ideas presented here have many important implications. I can only mention two implications briefly here. First, they suggest possible explanations for the large species differences between rats, primates and humans. It may be that the human neocortex must learn much more slowly than in less advanced animals, so that humans can take advantage of the much more complex, culturally transmitted, structure present in their environment. Also, it may be noted that the optimal procedure for learning structured environments may be to start with relatively large weight adjustments and then gradually reduce the size; in this way it is possible to guarantee convergence to the best possible set of weights (White, 1989; Darken and Moody, 1991). In this case, we would expect more rapid cortical learning early in life compared to later time periods. Second, these ideas provide one possible account of the phenomenon of infantile amnesia—the fact that we forget our early childhood experiences (Howe and Courage, 1993). According to the present theory, infantile amnesia could be due to the fact that the neocortical representations are changing relatively rapidly early in life, so that even if they are consolidated initially into the neocortical system they would tend to be overwritten by subsequent changes in the representations and connections. Later in life, as the neocortical representations became relatively stabilized, consolidated knowledge would tend to become less susceptible to interference.

Acknowledgments

The work reported here was supported by Grants MH00385 and MH47566, 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. This article is adapted from McClelland (1994).

References

Abraham, W. C. and Otani, S. (1991). Macromolecules and the maintenance of long-term potentiation. In: *Kindling and Synaptic Plasticity*, pp. 92–109, Morrell, F. (ed.). Birkhauser: Boston, MA.

Barnes, C. A. (1979). Memory deficits associated with senescence: A neurophysiological and behavioral study in the rat. *J. Comp. Physiol. Psychol.* **93,** 74–104.

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

Howe, M. L. and Courage, M. L. (1993). On resolving the enigma of infantile amnesia. *Psychol. Bull.* **113,** 305–326.

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

McClelland, J. L. (1994). Computational insights into the organization of human memory. In *Synapse-94*.

McClelland, J. L., McNaughton, B. L. and 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. *Psychol. Rev.* **102,** 419–457.

McCloskey, M. and Cohen, N. J. (1989). Catastrophic interference in connectionist networks: the sequential learning problem. In: *The Psychology of Learning and Motivation*, Vol. 24, pp. 109–165, Bower, G. H. (ed.) Academic Press: New York.

Plaut, D. C. and McClelland, J. L. (1993). Generalization with componential attractors: Word and nonword reading in an attractor network. *Proceedings of the 15th Annual Conference of the Cognitive Science Society*, pp. 824–829. Lawrence Erlbaum: Hillsdale, NJ.

Rumelhart, D. E. (1990). Brain style computation: learning and generalization. In: *An Introduction to Neural and Electronic networks*, pp. 405–420, Zornetzer, S. F., Davis, J. L. and Lau, C. (eds). Academic Press: San Diego, CA.

Rumelhart, D. E. and Todd, P. M. (1993). Learning and connectionist representations. In: *Attention and Performance XIV: Synergies in Experimental Psychology, Artificial Intelligence, and Cognitive Neuroscience*, pp. 3–30, Meyer, D. E. and Kornblum, S. (eds). MIT Press: Cambridge, MA.

Scoville, W. B. and Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *J. Neurol., Neurosurg. Psychiat.* **20,** 11–21.

Squire, L. R. and Cohen, N. (1979). Memory and amnesia: resistance to disruption develops for years after learning. *Behav. Neural Biol.* **25,** 115–125.

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

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

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