CHAPTER 25

Amnesia and Distributed Memory

J. L. McCLELLAND and D. E. RUMELHART

In several chapters in this book, we have argued for distributed models of learning and memory. In most of these cases, we have considered primarily *psychological* and *computational* reasons to prefer distributed models. In this chapter, we ask, can distributed models shed any light on the *biological* basis of memory? One possible answer would be "no"—we could fall back on the claim that distributed models are abstract descriptions, not concrete descriptions of the physiology of memory. Indeed, many of the specific distributed models we have considered in this book are somewhat "unphysiological" in several of their details. But the general idea of distributed memory (at least, within localized regions of the brain, as discussed in Chapter 3) does seem sufficiently consistent with what we know about the brain that the hypothesis that memory is physiologically distributed seems worth considering.

In this chapter, we consider this hypothesis in light of the phenomenon of *bitemporal amnesia*—the deficit in memory that is produced by a bilateral insult to the medial temporal lobes of the brain. Bitemporal amnesia is interesting from the point of view of distributed models because two distinct aspects of the phenomenon seem to suggest very different things about the biological plausibility of distributed models.

One prominent aspect of bitemporal amnesia is that it produces a *retrograde amnesia* that is *temporally graded.* After the precipitating insult, the individual is unable to remember recent events, but memory

for remote information appears to be intact. If there is recovery, as there is in many cases of bitemporal amnesia, much of the recent information that had been lost will return.

These aspects of amnesia seem to contradict the most basic assumptions of a distributed, superpositional model of memory. These models hold that all memories, old and new, are stored in the same set of connections. If this is so, why is it that an amnesic insult selectively disturbs the newer memories? And why is it that the memories that at first seemed to be lost can later be retrieved? The phenomenon seems to beg for an interpretation in which what is lost is access to that part of the memory store in which recent memories are held, rather than one in which all memories are superimposed in the same set of connections.

On the other hand, another prominent aspect of bitemporal amnesia seems to be highly consistent with a distributed model. Bitemporal amnesia produces a profound *anterograde* amnesia, as well as a *retrograde* deficit. That is, after the amnesic insult there may be a profound deficit in the ability to acquire new information. This is particularly true when amnesics are tested for their ability to recall or recognize specific individual events to which they have been exposed since onset of the amnesia. However, the amnesic deficit is not so profound, even in the severest cases, that the patient is unable to learn from repeated experience. For example, H. M., an extremely profound amnesic, is quite aware of his deficit, presumably as a result of repeatedly having been confronted with it. Milner (1966) reports that he often greets people by apologizing for not recognizing them, giving his memory deficit as his excuse. He remembers that he cannot remember, even though he cannot remember any particular occasion when he failed to remember.

This aspect of amnesia is quite naturally and directly accounted for by distributed models. We need only assume that the amnesic insult has resulted in a reduction in the size of the changes that can be made to connection strengths in response to any given event. Smaller changes will result in very weak traces of each individual episode or event, but, over repeated trials, what is common to a number of experiences will be gradually learned.

In summary, we appear to be faced by a paradoxical situation. One prominent aspect of bitemporal amnesia appears to argue against distributed models, while another appears to argue in favor of them.

In this chapter, we confront this paradox. First, we consider in more detail many of the basic aspects of retrograde amnesia. Then, we propose a model that appears to be capable of accounting for these facts within the context of a distributed model of memory. Several simulations are presented illustrating how the model accounts for various aspects of the empirical data on bitemporal amnesia, including the

temporally graded nature of retrograde amnesia and the ability to extract what is common from a set of related experiences. In a final section of the chapter, we consider some recent evidence suggesting that for certain kinds of tasks, amnesics show absolutely no deficits.

Basic Aspects of Amnesia

The term bitemporal amnesia was introduced by Squire (1982) to refer to the syndrome that is produced by a number of different kinds of insults that affect the medial portions of the temporal lobes in both hemispheres of the brain. The syndrome may be produced by bilateral electroconvulsive therapy (still widely in use as a treatment for severe depression), bilateral removal of the medial portions of the temporal lobes (as in patient H. M.), head trauma, or in several other ways. The syndrome is marked by the following characteristics (see Squire, 1982, for a more detailed discussion):

- The anterograde and retrograde amnesias produced by the insult appear to be correlated in extent. While there are some reports of dissociation of these two aspects of amnesia, it is well established in cases of amnesia due to electroconvulsive therapy that anterograde and retrograde amnesia are correlated in severity; both develop gradually through repeated bouts of electroconvulsive therapy.

- The anterograde amnesia consists of a deficit in the acquisition of new knowledge accessible to verbal report or other explicit indications that the subject is aware of any particular prior experience; somewhat more controversial, it also consists of a more rapid loss of information once it has been acquired to a level equal to normal levels of acquisition through repeated exposure.

- The retrograde amnesia consists of an inability to give evidence of access to previous experiences within a graded temporal window extending back over an extended period of time prior to the amnesic insult. The size of the window varies with the severity of the amnesia, and good evidence places it at up to three year's duration based on careful experimental tests.

- Most strikingly, memories that appear to be lost after an amnesic insult are often later recovered. As the ability to acquire new memories returns, so does the ability to remember old ones that had previously been lost. The recovery is gradual, and it is as if the temporal window of retrograde amnesia shrinks. There is generally a residual, permanent amnesia for events surrounding the insult that caused the amnesia, extending variously from minutes to days from the event.

A Resolution to the Paradox

As we have already noted, the temporally graded nature of the retrograde aspect of bitemporal amnesia appears to suggest that recent memories are stored separately from older ones. However, it is possible to account for this aspect of the phenomenon in the context of a distributed model if we make the following assumptions. First, we assume that each processing experience results in chemical/structural change in a large number of connections in which many other traces are also stored, but that each new change undergoes a gradual consolidation process, as well as a natural decay or return to the prechange state. Thus, the changes resulting from a particular experience are widely distributed at one level of analysis, but at a very fine grain, within each individual connection, each change in its efficacy has a separate consolidation history.[1] Second, we assume that consolidation has two effects on the residual part of the change: (a) It makes it less susceptible to decay; and (b) it makes it less susceptible to disruption. These assumptions can explain not only the findings on the temporally graded nature of retrograde amnesia, but also the fact that memory appears to decay more rapidly at first and later decays more slowly.

So far this explanation simply takes existing consolidation accounts of the amnesic syndrome (e.g., Milner, 1966) and stipulates that the changes are occurring in synapses that they share with other changes occurring at other points in time. However, we need to go beyond this account to explain two of the important characteristics of the bitemporal amnesic syndrome. First, the hypothesis as laid out so far does

[1] When we speak of connections between units, even if we think of those units as *neurons*, we still prefer to use the term *connection* somewhat abstractly; in particular, we do not wish to identify the connection between two units as a *single* synapse. Two neurons may have a number of different physical synapses. The total strength of these synapses determines the strength of the connection between them.

not explain recovery; second, it does not explain the coupling of anterograde and retrograde amnesia.

To capture these two important aspects of the syndrome, we propose that there exists a factor we call γ (*gamma*) that is depleted by insult to the medial temporal lobes. Gamma serves two functions in our model: (a) it is necessary for consolidation; without γ, new memory traces do not consolidate; and (b) it is necessary for expression; without γ, recent changes in the connection between two units do not alter the efficacy of the connection; they are just ineffectual addenda, rather than effective pieces of new machinery. Implicit in these assumptions is a third key point that γ is only necessary during consolidation. Fully consolidated memories no longer need it for expression.

Some Hypothetical Neurochemistry

To make these ideas concrete, we have formulated the following hypothetical account of the neurochemistry of synaptic change. While the account is somewhat oversimplified, it is basically consistent with present knowledge of the neurochemistry of synaptic transmission, though it should be said that there are a number of other ways in which connection strengths could be modulated besides the one we suggest here (for an introductory discussion of current understanding of synaptic function and synaptic modification, see Kandel & Schwartz, 1981).

The account goes as follows. The change to the connection from one unit to another involves adding new receptors to the postsynaptic membrane (the one on the input unit) (see Figure 1). We assume that both

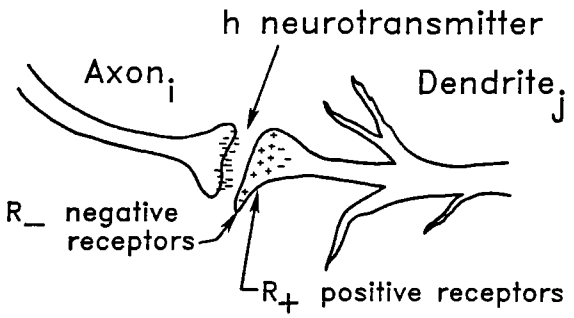

FIGURE 1. A connection between two units, as we conceptualize it in the amnesia model. Note that both positive and negative changes involve addition of new receptors. See text for discussion.

positive and negative changes involve the addition of receptors; in both cases, there must be new structure to consolidate for the model to work properly. In the figure, we have drawn the connection between two units as though it occurred at a single synapse and was not mediated by interneurons, though neither of these assumptions is excluded by the quantitative structure of the model.[2]

A cartoon of one of the receptors is shown in Figure 2. Receptors are, of course, known to be the physical structures whereby neurotransmitters released by the presynaptic neuron influence the potential of the postsynaptic neuron. To be functional, though, our hypothetical receptors must be clamped in place at each of several γ-binding sites by molecules of γ—this is the aspect of the model that is the most speculative. The probability that a site is bound depends, in turn, on the

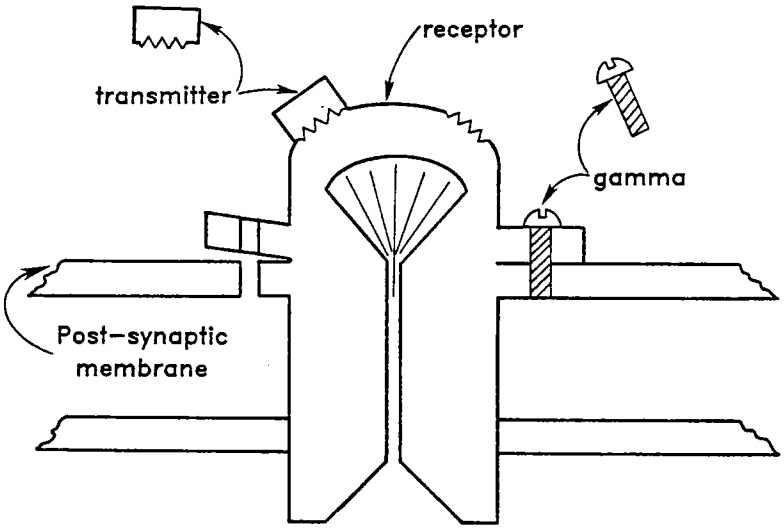

FIGURE 2. A cartoon of a receptor, showing its location in the postsynaptic membrane and illustrating the role of the transmitter substance, and of the hypothetical substance γ, which acts to bind the receptor into the membrane.

[2] As pointed out in Chapter 20, real neurons are generally thought to have either excitatory or inhibitory connections but not both. Our model could be brought into line with this idea if we assumed that negative (inhibitory) connections between two units actually occurred at excitatory synapses onto inhibitory interneurons, rather than on direct connections between two neurons. Connections onto these inhibitory interneurons would have to be trained, of course, using something like the generalized delta rule (Chapter 8). This revision of our assumptions would increase the complexity of the model but would not change its basic properties; therefore we have retained the less realistic assumption that positive and negative increments can be stored in the same connections.

concentration of γ in the environment of the synapse. In this model, consolidation amounts to the "hardening" or fixation of the γ-binding sites, while they are occupied by a molecule of γ. Thus, consolidation can only occur at bound sites. Consolidation is a process like the setting of glue, but it is thought to be probabilistic and all-or-none at each site rather than continuous.

As we have already seen, γ is essential for consolidation. In addition, we assume that it is necessary for the receptor to function. Once a site is consolidated, however, γ is irrelevant to it, just as a clamp is irrelevant once a glue-joint is set. Thus, unconsolidated sites depend on γ, but consolidated ones do not.

On this view, bitemporal amnesia simply amounts to taking away the clamps. Old, fully consolidated synaptic changes no longer require them, and new ones cannot function without them and will decay without becoming consolidated. But what of memories in an intermediate stage of consolidation? Here, we assume the consolidation process has gone far enough so that the structures will not break up rapidly without γ, but that it has not gone so far that they actually function effectively without it. When γ returns, after a period for recovery, they may still be there, so they will be able to function again and even continue to consolidate.

A Quantitative Formulation of the Model

Let us now formalize these assumptions in a quantitative model. We assume that time is broken up into a number of discrete ticks. In the simulations each tick represents about an hour of real time. On each tick, an unconsolidated site is bound by γ with a probability ρ, given by the *law of mass action* (this law governs a large number of chemical and biochemical processes):

$$\rho = \frac{\gamma}{1 - \gamma}.$$

This equation has the property that at high concentrations of γ (much greater than 1), all unconsolidated sites will be bound, but at low concentrations (less than about .2), the probability of being bound is roughly linear with γ.

Now, in each tick, an unconsolidated site may become consolidated or "fixed" with some probability f, but only if it is bound. Thus, the probability of consolidation of unbound site i per tick is just

$$p_c(site_i) = f\rho.$$

For a receptor to be *functional* at a particular tick, *all* its sites must be either consolidated or bound with γ. Each unconsolidated site is assumed to be independent of the others, so the probability that receptor i will be active, $p_a\,(receptor_i)$, is just

$$p_a\,(receptor_i) = \rho^u$$

where u is simply the number of unconsolidated sites.

Finally, receptors may be lost from the postsynaptic membrane. Each site contributes multiplicatively to the probability that the receptor will be lost. That is, the probability that receptor i will be lost is simply the *product* of the susceptibilities for each site. The susceptibility of consolidated sites, θ_c, is assumed to be small enough so that for completely consolidated receptors the probability of loss is very very small per tick; though over the course of years these small probabilities eventually add up. The susceptibility of unconsolidated sites, θ_u, is relatively large. For any given receptor, some number c of its sites are consolidated at any given time and u sites are not. The probability of receptor loss per tick, $p_l\,(receptor_i)$ simply becomes

$$p_l\,(receptor_i) = (\theta_c)^c\,(\theta_u)^u.$$

Relation to Other Accounts of Amnesia

Most attempts to account for temporally graded retrograde amnesia quite naturally involve some form of consolidation hypothesis, and our model is no exception to this. However, other accounts either leave the nature of the consolidation process unspecified (e.g., Milner, 1966) or give it some special status. For Wickelgren (1979), who has the most concretely specified account of retrograde amnesia, memory trace formation involves a "chunking" or unitization process whereby each memory trace is organized under its own superordinate or "chunk" unit. A number of other authors have proposed accounts with a similar flavor (e.g., Squire, N. J. Cohen, & Nadel, 1984).

In keeping with the view that our model can be implemented in a distributed memory system, our model of consolidation does not involve anything like chunking of a memory trace under a single superordinate unit. Instead, it simply involves the fixation of memory traces in a time-dependent fashion, dependent only on a single, global factor: the concentration of γ.

This difference means that our model gives the hippocampus a rather different role than it is taken to have in other theories. Theorists

generally have not imagined that the hippocampus is the actual site of memory storage, for on that view, it would be difficult to explain why retrograde amnesia is temporally graded, unless only recent memories are thought to be stored there. But the hippocampus is often thought to play a very important role in memory trace formation. To Wickelgren, for example, the hippocampus is the organ of unitization—it is the units in the hippocampus that bind the pieces of a memory trace together into chunks. In our model, we imagine that the primary role of the hippocampus in memory formation is to produce and distribute γ to the actual memory storage sites. This does not mean that we believe that this is the *only* function of the hippocampus. An organ as complex as the hippocampus may well play important information processing roles. However, as we shall see, this role is sufficient to provide quite a close account of a number of aspects of the amnesic syndrome.

Simulations

The primary goal of the simulations was to demonstrate that, with the simple assumptions given above, we could account for the main aspects of the coupled phenomena of anterograde and retrograde amnesia, using a single set of values for all of the parameters of the model, only allowing γ to vary with the assumed amnesic state of the subject. Since the phenomena range over a wide range of time scales (hours or even minutes to years), this is by no means a trivial matter.

Rather than embedding the assumptions about amnesia in a full-scale distributed model, we have simply computed, from the above assumptions, what the residual fraction (or residual *functional* fraction) of the memory trace would be at various times after the learning event, under various conditions. The assumption here, basically, is that each memory trace is made up of a large number of receptors distributed widely over a large number of connections. The fraction of the total that remains and is functional at any particular time gives the "strength" of the memory trace. To relate the results of these simulations to data, it is sufficient to assume that the size of the residual functional fraction of a memory trace is monotonically related to accuracy of memory task performance.

Of course, memory task performance, and indeed the effective residual strength of a memory trace, does not depend only on the hypothetical biochemical processes we are discussing here. For one thing, there is interference: New memory traces acquired between a learning event and test can change connection strengths in such as way as to actually reverse some or all of the changes that were made at the time of the

original encoding event, producing what were classically known as retroactive interference effects. There will be proactive interference effects as well in a distributed model (see Chapter 3). Additionally, as time goes by, there will be changes in the mental context in which retrieval takes place; all of these factors will contribute to the apparent strength of a memory trace as observed in experiments. The point of our model of amnesia is not to deny the importance of such factors; we simply assume that performance in a memory task varies with the residual functional fraction of the original trace, all else being equal.

The values of the parameters are shown in Table 1. The significance of the particular values chosen will become clear as we proceed.

Anterograde amnesia: Smaller functional fractions at all delays. As we have said, our model assumes that amnesia amounts to a reduction in the size of γ. Reducing the size of γ does not reduce the size of the memory trace—the number of receptors added—but it does greatly reduce their effectiveness: For a receptor to be functional, *all* of its sites must be either bound with gamma or consolidated. Initially, before any consolidation has occurred, the probability that a receptor will be functional, or active, is

$$p_a\,(receptor_i) = \rho^n$$

where n is the number of sites on the receptor. For normals, we take ρ (that is, $\gamma/(1-\gamma)$) to be .5. Since each receptor has three sites, p_a will be $.5^3$ or .125; if amnesia reduces the concentration of γ by a factor of 9, the effect will be to reduce ρ to .1, and p_a to .001. In general, with n_{sites} equal to 3, reducing ρ by a particular fraction of the normal

TABLE 1

PARAMETERS USED IN SIMULATIONS OF THE AMNESIA MODEL

Parameter Name	Value
n_{sites}	3
f	.00007
θ_c	.02
θ_u	.25
γ_{Normal}	1.0

Note: Rate parameters f, θ_c, and θ_u are given on a per-hour basis.

value will reduce the effective strength of the initial memory trace by the *cube* of that fraction.

Correlation of retrograde and anterograde amnesia. The model produces retrograde amnesia, as well as anterograde amnesia, for the unconsolidated portion of a memory trace. The reason for this is that the *expression* of unconsolidated memories depends on γ; thus, it applies both to posttraumatic memories and to memories formed before the trauma. Indeed, the severity of anterograde and retrograde amnesia are perforce correlated in the model since both depend on γ for the expression of traces that have not had time to consolidate.

Retrograde amnesia: Older traces are less dependent on gamma. One of the most interesting aspects of retrograde amnesia is the fact that it is temporally graded. Indeed, in the data collected by Squire, Slater, and Chace (1975), shown in Figure 3, it is not only graded, but bitemporal amnesics actually show worse memory for recent events than for those about three to five years old. This matches the clinical impression for patients such as H.M., of whom it is reported that his retrograde amnesia initially extended over a period of one or two years

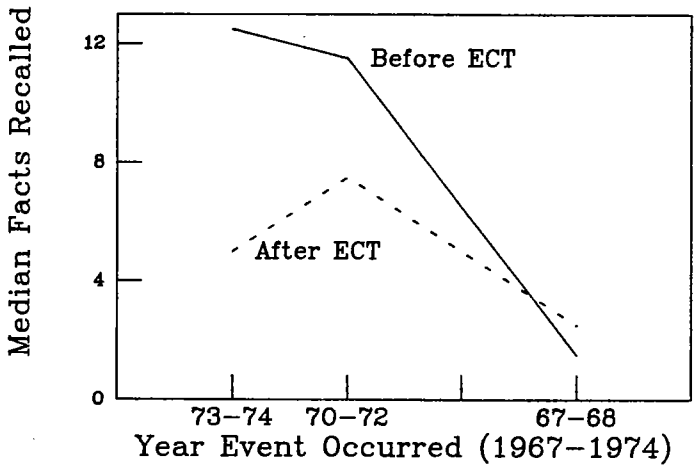

FIGURE 3. The striking temporally graded retrograde amnesia observed in patients whose amnesia was induced by electroconvulsive therapy. Patients served as their own controls, based on an alternate form of the test given prior to the beginning of treatment. (From "Retrograde Amnesia: Temporal Gradient in Very Long-Term Memory Following Electroconvulsive Therapy" by L. R. Squire, P. C. Slater, and P. Chace, 1975, *Science*, *187*, Copyright 1975 by the American Association for the Advancement of Science. Reprinted by permission.)

(Milner, 1966), and from victims of head trauma (Russell & Nathan, 1946). This pattern has been replicated many times, and the tests used by Squire et al. rule out artifacts that have plagued clinical assessments of the severity of retrograde amnesia. This inverted U-shaped curve for the relation between age of memory and memory test performance provides quite a challenge to theories of retrograde amnesia. However, this effect is a natural consequence of our model since old memories, though based on smaller residual traces than newer ones, are less dependent on γ for their expression. Indeed, when a receptor reaches a point where all of its sites are consolidated, it no longer depends on γ at all.

A simulation capturing the essential features of temporally graded retrograde amnesia as represented in Squire et al. is shown in Figure 4. The simulation produces a continual erosion in functional strength for normals which is almost linear against the log of time over the range of times covered by the simulation. In contrast, for amnesics, the function is decidedly (inverted) U-shaped: Functional trace strength reaches a peak at about 2 to 3 years with these parameters and then falls of gradually thereafter, following the same trajectory as the strength of the trace for normals. The location of the peak in the RA function depends on all of the parameters of the model, but the primary ones are the consolidation rate parameter f, which is .00007, and the rate of decay from consolidated memory, which is $(\theta_c)^n$, or

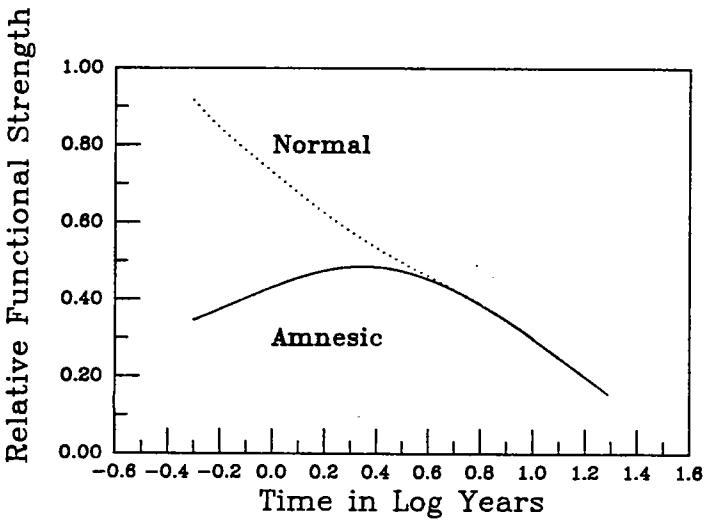

FIGURE 4. Simulation of temporally graded retrograde amnesia. Effective trace strength as a function of time is years preceding sudden onset of amnesia. Effective trace strength is normalized so that a value of 1.0 corresponds to the normal value at five months.

8×10^{-6}. It should be noted that these figures are *per hour*. The consolidation parameter translates into a consolidation rate of about 50% per year, per receptor site. The average time it takes for all of the sites on a receptor to become consolidated is longer than this, of course, but only by about a factor of 2 for the case of $n_{sites} = 3$; this is essentially the factor that determines where the curve for amnesics will catch up with the curve for the normals. The decay rate from fully consolidated memory, which translates into about 7% per year or 50% per decade, essentially determines the overall slope of the normal function and the tail of the amnesic function.

Recovery of lost memories: The return of the partially consolidated trace. Perhaps even more interesting than the fact that retrograde amnesia is temporally graded is the fact that it recovers as the ability to acquire new memories recovers. In the case of retrograde amnesia induced by electroconvulsive therapy, Squire, Slater, and Miller (1981) showed that the severe retrograde amnesia for pretreatment memories recovers over the course of several months, at the end of which test performance is back to pretreatment levels. In our model, since retrograde amnesia is due to the fact that loss of γ renders traces ineffective, it is not surprising that the return of γ will render them effective again. However, the phenomenon is somewhat more subtle than this, for recovery is not generally thought to be complete. There is usually some loss of memory for events in the time period preceding the onset of the amnesia, and the precipitating event is almost never recalled; this is particularly striking in head trauma patients, who often do not know what hit them, even if they had seen it at the time (Russell & Nathan, 1946).

To examine how well our model can do at reproducing these aspects of amnesia, we ran the model in the following simulated amnesia-recovery experiment. The model was made amnesic at some time t_a and was left in this state until some time t_r, at which point we assumed recovery occurred. Of course, real recovery is gradual, but for simplicity, we assumed that it was a discrete event. We then asked what fraction of a trace laid down before the onset of amnesia remained, relative to the fraction that would have remained had there been no insult. The results of the simulation are shown in Figure 5. Each curve shows the strength of the recovered trace, relative to the strength it would have had with no intervening amnesia, as a function of the duration in months of the amnesic episode. Clearly, for memories a year or more old, the trace recovers to nearly premorbid levels at the end of the amnesic episode, even if it lasts as long as a year. For memories laid down within the day of the the amnesic event, however, the bulk of the trace is gone by the end of the amnesic episode, even if it lasts only

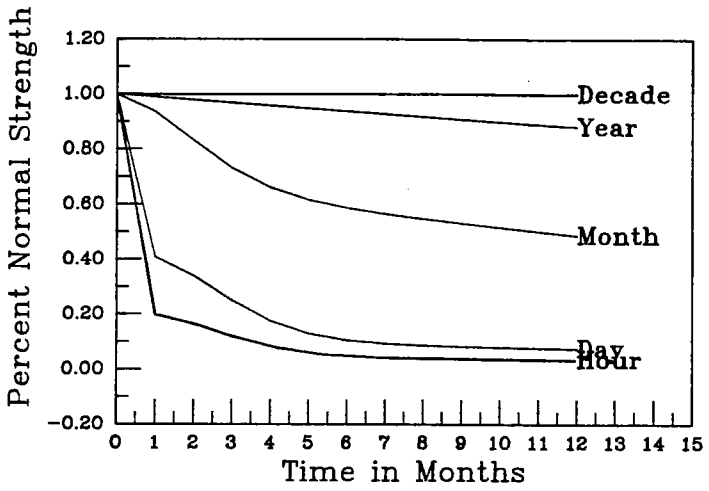

FIGURE 5. Simulated recovery of premorbid memories as a function of time in the amnesic state and age of memory at onset of the amnesia. The age ranges from an hour to a decade, as indicated by the labels on the curves.

a month. Memories that are a month old at the onset of the amnesia show an intermediate pattern. If the amnesia is relatively brief, they survive quite well; but if it lasts several months, they weaken considerably, relative to the strength they would have had in the absence of amnesia.

The loss of memory trace strength during amnesia is a result of the great reduction in the opportunity for consolidation during the amnesic interval. We now turn to a more direct consideration of this matter.

Do amnesics forget faster than normals? A number of studies (Huppert & Piercy, 1978; Squire, 1981) have reported that bitemporal amnesic subjects appear to forget more rapidly than normals, even if equated with normals for the amount of initial learning. The effect is generally rather small, and it is controversial because the equating of groups on initial performance requires giving amnesics considerably more training than normals, over a longer period of time. These differences could possibly change the basis of learning and other qualitative aspects of the task as experienced by amnesic and normal subjects. It is interesting, then, to consider whether our model would predict such a difference.

The model does predict a difference in rate of trace decay between amnesic and normal subjects. Though γ does not influence the rate of trace decay directly, it does influence the rate of consolidation, and

consolidation drastically influences the rate of decay. Completely unconsolidated traces decay at a rate of $.25^3 = 1.5\%$ per hour, or about 30% per day, and are reduced to 1/100,000 of their initial strength in a month. Consolidated traces, on the other hand, decay at a rate of only $.02^3 = .0008\%$ per hour, or less than 1% per month. As each site becomes fixed, it retards decay by a factor of 12. Thus, to the extent that consolidation is occurring, memory traces are being protected against rapid loss; without any consolidation, trace strength falls precipitously. An illustration of this effect is shown in Figure 6. At higher values of γ, the drop of trace strength decelerates much earlier than at lower values of γ, leaving a much larger residual trace.

Unfortunately, the effect shown in Figure 6 does not happen in the right time scale to account for the differential decay of normal and amnesic memory over hours, as reported by Huppert and Piercy (1978) and Squire (1981). In fact, the effect does not really begin to show up until after about 10 days. The reason is clear: The consolidation rate is so slow (.0007/hour, or 5% per month) that very little consolidation happens in the first month. Thus, it appears that the value of the consolidation rate parameter required to account for temporally graded retrograde amnesia and U-shaped memory performance on a time scale of years is much too slow to account for differences in trace decay on a time scale of hours.

One possible reaction to this state of affairs would be to search for alternative interpretations of the apparent differences between normals

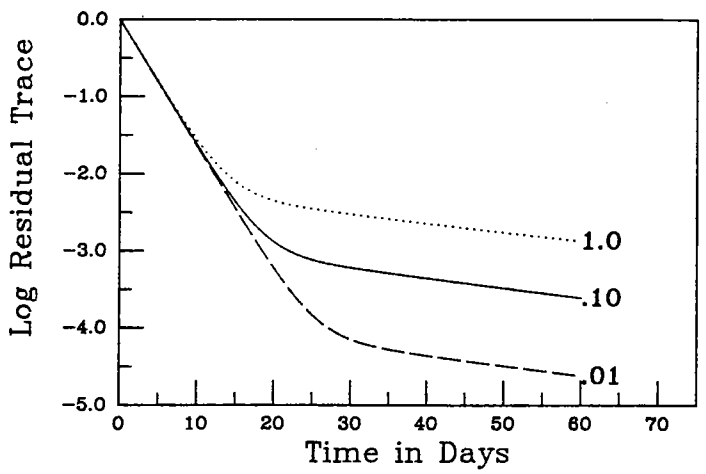

FIGURE 6. Log of the residual strength of a trace acquired at Day 0, as a function of the concentration of γ. The values of 1.0 and 0.1 correspond to the values described as normal and amnesic in the text.

and amnesics in day-scale decay rates. However, there are other reasons to believe that there is consolidation on a shorter time scale than we get with our model and the parameters in Table 1. For one thing, a single bout of electroconvulsive therapy that produces a brief and mild amnesia nevertheless appears to produce permanent loss of memory for the shock treatment itself. Such an effect seems to suggest that there is consolidation going on over a shorter time-scale.

The model might capture all of the data if we assumed that there are two separate phases to consolidation, both of them dependent on γ: one that occurs on a relatively short time scale and is responsible for the differences in day-scale decay rates, and one that occurs on a very long time scale and is responsible for extended temporally graded retrograde amnesia. As things stand now, traces decay rather slowly over hours, but, over the course of a month, they are reduced to about half of one percent of their original strength. Though we do not know exactly how to scale trace strength against response probability, it seems likely that we forget more quickly over hours but more slowly over days and months than in the present version of the model.

Summary. The model provides, we think, an appealing, unified account of most aspects of anterograde and retrograde amnesia, simply by assuming that the amnesic insult depletes γ and that recovery amounts to its gradual return to pretraumatic levels. By adding an additional stage of consolidation, the model could be made to span the very wide range of time scales, ranging from hours to years, of the coupled phenomena of anterograde and retrograde amnesia as they appear in the bitemporal amnesic syndrome.

Most importantly, the model shows clearly that there is no incompatibility between the phenomenon of temporally graded retrograde amnesia and distributed representation. So far, however, our account of amnesia has not really depended on the features of our distributed model. In the next section we will consider aspects of the amnesic syndrome which do seem to point toward distributed models.

RESIDUAL LEARNING AND SPARED LEARNING IN BITEMPORAL AMNESIA

As we noted briefly before, there are some domains in which amnesics exhibit what are generally described as *spared learning* effects: They show no noticeable deficits when compared to normal subjects. There is now a very large literature on these spared learning effects.

The following summary seems to capture the basic characteristics of what is spared and what is not.

While amnesics seem to be highly deficient in the ability to form accessible traces of particular individual episodic experiences, they seem to be completely spared in their ability to learn certain types of skills that require no explicit access to the previous processing episodes in which the skill was acquired (N. J. Cohen, 1981; N. J. Cohen, Eichenbaum, Deacedo, & Corkin, 1985). In addition, they show apparently normal repetition priming effects in experiments involving such tasks as perceptual identification, in which the subject must simply identify a briefly flashed word in a short exposure (see Schacter, 1985, for a review). These effects may be strongly and strikingly dissociated from the subjects' verbally expressed recollections. Thus, H. M. has acquired a skill that allows him to perform perfectly in solving the Tower of Hanoi problem, without becoming aware that he has actually ever performed the task before and without knowing (in a conscious, reportable sense) even what constitutes a legal move in the Tower Puzzle (N. J. Cohen et al., 1985). Also, amnesic subjects show normal effects of prior exposure to words in perceptual identification and related tasks, without necessarily having any awareness of having seen the words or even participating in the priming portion of the task. Between these two extremes lies a gray zone. Within the domains where learning is impaired, even the densest amnesics seem to learn, however gradually, from repeated experience (Schacter, 1985). First, we will consider these *residual learning* effects from the point of view of distributed memory. Then, we will examine the more striking *spared learning* effects.

Residual Learning in Bitemporal Amnesia

As we noted early in this chapter, distributed models provide a natural way of explaining why there should be residual ability to learn gradually from repeated experience within those domains where amnesics are grossly deficient in their memory for particular episodic experiences. For if we imagine that the effective size of the increments to the changes in synaptic connections is reduced in amnesics, then the basic properties of distributed models—the fact that they automatically extract the central tendency from a set of similar experiences and build up a trace of the prototype from a series of exemplars—automatically provides an account of the gradual accumulation of knowledge from repeated experience, even in the face of a profound deficit in remembering any specific episode in which that information was

presented. Distributed models are naturally incremental learning models, and thus they provide a very nice account of how learning could occur through the gradual accumulation of small traces.

We call the hypothesis that anterograde amnesia amounts to reducing the effective size of the increments the *limited increment hypothesis*. For bitemporal amnesics, the effective size of the increments is limited by the depletion of γ; in other forms of amnesia (which also show similar kinds of residual learning) the size of the increment might be limited in other ways. According to the limited increment hypothesis, residual learning is simply a matter of the gradual accumulation of information through the superimposition of small increments to the connection strengths.

To illustrate this point, we have carried out a simulation analog of the following experiment by N. J. Cohen (1981). Amnesic subjects and normal controls were seated in front of an apparatus with a movable lever. On each trial of the experiment, the subject was asked to move the lever until it reached a stop set by the experimenter. The experimenter then moved the lever back to the start position and removed the stop. After a variable delay, the subjects were asked to reproduce the previous movement. Such trials are referred to as *reproduction trials*.

At the end of each group of three trials, subjects were asked to reproduce their impression of the average distance they had been asked to move the lever, based on all of the preceding trials in the experiment. Such trials will be called *averaging trials*.

The results of the reproduction task were as expected from the fact that amnesics have very poor memory for specific experiences; at very short delay intervals, amnesics did as well as normals, but at longer intervals, they were grossly impaired, as measured by the deviation of the reproduced movement from the training movement (Figure 7). However, amnesics did no worse than normals at reproducing the average movement. The experiment was divided into four parts: In the first and last parts, the movements were all relatively long; and in the two intermediate parts, the movements were all relatively short (some subjects had the long and short trials in the other order). At the end of each block of trials, both groups accurately reproduced the average movement for that block (For the long blocks, movements averaged 42.6 degrees for the normals and 41.3 for the amnesics; for the short blocks, movements averaged 30.8 for the normals and 30.6 for the amnesics).

We simulated this phenomenon using the distributed memory model described in Chapter 17. Briefly, that model consists of a single module, or set of units, with each unit having a modifiable connection to each other unit. The units in the module receive inputs from other

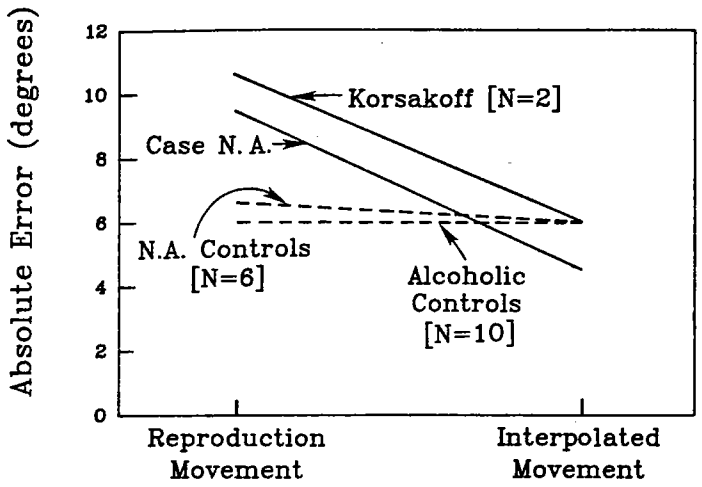

FIGURE 7. Accuracy of reproduction movements by amnesics and normal controls in the lever placement experiment described in text. (From "Neuropsychological Evidence for a Distinction Between Procedural and Declarative Knowledge in Human Memory and Amnesia" by N. J. Cohen, 1981, doctoral dissertation, University of California, San Diego. Copyright 1981 by N. J. Cohen. Reprinted by permission.)

units via the modifiable connections, as well as external inputs from stimulus patterns. Processing in the module begins with all units at a resting activation of 0 and the presentation of an external input pattern. In this case, the module consisted of 16 units, and each input pattern was a vector of 16 excitatory or inhibitory inputs. When a pattern is presented, it begins to drive the activations of the units up or down as a result of its direct effects; the units then begin to send excitatory and inhibitory signals to the other units via the modifiable connections. For patterns that have previously been stored in the connections among the units, the internal connections produce an enhancement of the pattern of activation over and above what would be produced by the external input alone; if, however, the external input is very dissimilar (orthogonal) to the patterns that have been stored in the module on previous trials, there will be little or no enhancement of the response.

On each trial of the experiment, a new distortion of the same 16-element prototype pattern was presented to the module, and connection strengths were adjusted after each trial according to the delta rule (see Chapter 17 for details). We then tested the module in two ways: First, to simulate Cohen's reproduction test, we looked at the magnitude of the model's response to the pattern it had just been shown. For the averaging test, we looked at the magnitude of the model's response to

the prototype. Note that these test trials were run with connection strength modification turned off, so each test was completely uncontaminated by the previous tests.

In keeping with the limited increment hypothesis, we assumed that the difference between amnesics and normals in Cohen's experiment could be accounted for simply by assuming that amnesics make smaller changes to the strengths of the connections on every learning trial. To show that the model shows residual learning of the prototype under these conditions, we ran the simulation several times, with three different levels of the increment strength parameter η from the equation for the delta rule, which we reproduce here:

$$\Delta w_{ij} = \eta \delta_i a_j .$$

The results of the simulation are shown in Figure 8. As the figure indicates, the larger the size of η, the more strongly the model responds to the immediately preceding distortion of the prototype. But, after a few trials, the response to the *central tendency* or prototype underlying each distortion is as good for small values of η as for larger ones. In fact, response to the prototype is actually *better* when the model is "amnesic" (low η) than when it is "normal" (high η); in the latter state, the connections are continually buffeted about by the latest distortion, and the model has trouble seeing, as it were, the forest for the trees.

In the figure, there is a gradual improvement in the response to the immediately preceding stimulus for small increment sizes. This occurs only because the stimuli are all correlated with each other, being derived from the same prototype. For a sequence of unrelated stimuli, the response to each new input shows no improvement over trials.

This pattern of performance is very reminiscent of the pattern seen in several of the experiments performed by Olton and his colleagues (see Olton, 1984, for a review). They have trained rats to run in two different mazes, each having two choice points. At one of the choice points, the response was always the same for a given maze (we call this the maze-dependent choice); at the other choice point, the response that had to be made varied from trial to trial, based on the response the rat made on the preceding trial (we call this the trial-dependent choice). The principal finding of these experiments is that rats with hippocampal lesions show gross impairment in the ability to make the right trial-dependent choice, but show no impairment in the ability to make the right maze-dependent choice if they were trained up on the task before surgically induced amnesia. The acquisition of the maze-dependent choice is slowed in rats trained after surgery, but these animals eventually reach a point where they can perform as well as normals. Such animals show near chance performance in the trial-dependent choice

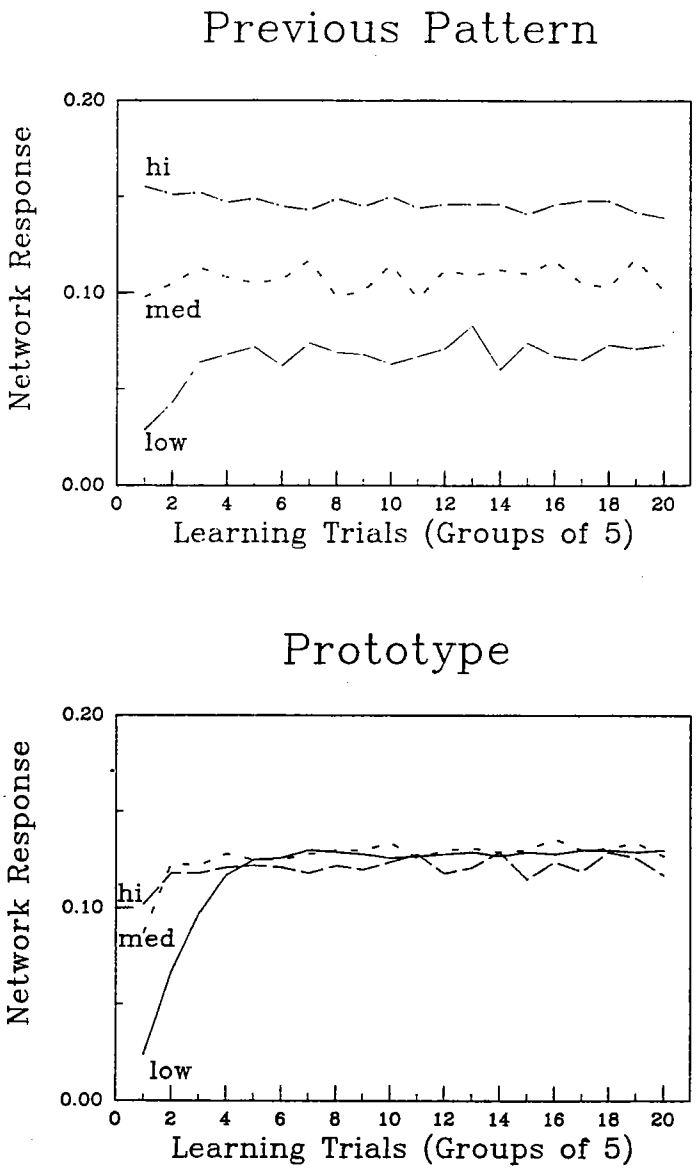

FIGURE 8. Network response for the most recent distortion of the prototype pattern and for the prototype pattern itself, as a function of test trials, at three different levels of the increment size parameter η. Network response is the dot product of the external input to each unit with the internal input generated by the network.

after surgery, even if they had already acquired the ability to do this part of the task before the surgery.

Such a pattern of results is completely consistent with the limited increment hypothesis: Performance in the trial-dependent choice requires, by the design of the task, that the subject rely on a memory trace of the preceding trial of the experiment, whereas performance on the maze-dependent choice can be based on a composite memory trace acquired gradually over repeated experience in the same maze. No separate mechanisms for retaining recent episodes, as opposed to more general memories, is required.

In summary, there are a variety of phenomena, both in human and animal amnesia, which fit in very well with the kind of gradual, residual learning we see in our distributed model. Distributed models naturally and automatically pull out what is common to a set of experiences, even if, or one might even say especially when, the traces of the individual experiences are weak.

It is worthwhile to note that this property of distributed models would not be shared by all learning models, especially those that rely on some mechanism that examines stored representations of specific events in order to formulate generalizations, as in the ACT* model (J. R. Anderson, 1983), or in Winston's (1975) approach to learning. For on such models, if the individual traces are impaired, we would expect the generalization process to be impaired as well. Of course, such models could account for these effects by assuming that each episode is stored in two different ways, once for the purpose of learning generalizations and once for the purpose of remembering the details of particular experiences. Thankfully, our distributed model does not require us to duplicate memory stores in this way; residual learning based on small increments drops out of the basic superpositional character of the model.

Spared Learning of Skills

More striking that these residual learning effects is the phenomenon of spared learning: The fact that the acquisition of a variety of general skills occurs at roughly the same rate in normal and amnesic subjects. This fact has been taken as evidence that the brain maintains a distinction between those structures underlying explicitly accessible episodic and semantic information on the one hand and those underlying general cognitive skills on the other (N. J. Cohen et al., 1985).

While this conclusion is certainly plausible, it is worth noting that there are other possibilities. One that we have considered is the possibility that limited increments to connection strengths make a difference for some kinds of tasks but not for others. The simulations reported in

the previous section indicated that this can sometimes be the case; in fact, they indicated that, as far as extracting the prototype or central tendency of an ensemble of experiences is concerned, it can sometimes be better to make smaller changes in connection strengths.

The preserved *skill learning* observed in many tasks appears to be the kind of learning that may be relatively unaffected by the size of the changes made to connections. For we can view skill learning as the process of learning to respond to *new* stimuli in a domain, based on experience with previous examples. For example, consider the mirror-reading experiment of N. J. Cohen and Squire (1980). In this experiment, subjects were required to read words displayed reflected in a mirror so that they had to be read from right to left. In this task, both amnesic and normal subjects learn gradually. Though normals learn to read specific repeated displays much more quickly than amnesics, both groups show equal transfer to novel stimuli.

To assess transfer performance in the simple distributed model described in the previous section, we observed the response of the model to new input patterns, after each learning trial. The results of the simulation are shown, for three levels of η, in Figure 9. Though there are initial differences as a function of η, these differences are considerably smaller than the ones we observe on reproducing old associations. And, at a fairly early point, performance converges, independently of the level of η. As with learning the prototype, there is a slight advantage for smaller values of η, in terms of asymptotic transfer performance, though this is difficult to see in the noise of the curves.

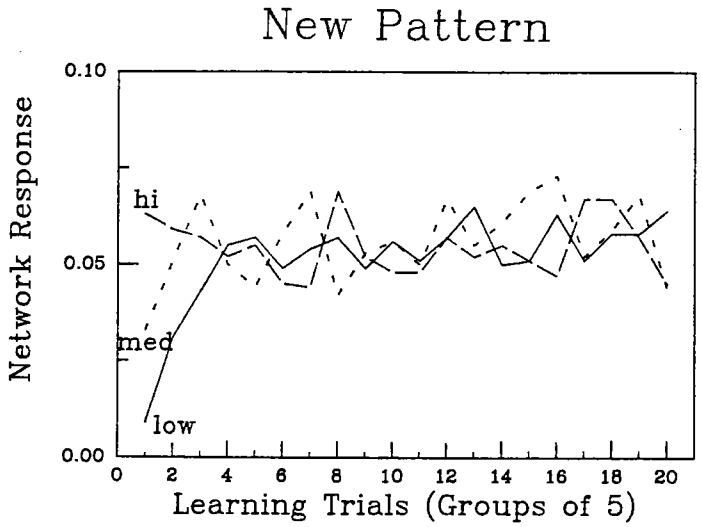

FIGURE 9. Network response to transfer patterns (new distortions of the prototype) for three levels of η, as a function of learning.

This simulation does not capture spared learning of skills perfectly, since our simulated normals approach asymptote more rapidly than our simulated amnesics. However, if such skills really consisted of many small skills, the differences might not be noticeable. We have not yet developed a version of our model in which there are no differences as a function of η. We still consider it an open question as to whether we will succeed. It may turn out that there are other distributed models (perhaps involving hidden units) in which rate of learning is quite insensitive to large differences in sizes of increments on certain kinds of measures for certain kinds of tasks. This is a matter we are continuing to pursue as our explorations of memory and learning continue.

Setting this possibility aside, let us suppose for the moment that episodic and semantic memory are learned in one memory system, dependent on medial temporal lobe structures, and general skills are learned in a different system. This view raises a question: Why should this be? Why should the brain make this distinction? We can actually provide one possible answer to this question based on our observations of the properties of the simple distributed model presented in the simulations. These observations suggest that large changes in connection strengths may be better for storing specific experiences but may do more harm than good for gradually homing in on a generalizable set of connection strengths.[3]

On the basis of these observations, we might propose that the temporal lobe structures responsible for bitemporal amnesia provide a mechanism that allows large changes to connections to be made in parts of the system in which memories for specific experiences are stored, but other parts of the cognitive system make use of a different mechanism for changing connections that results in the smaller changes that are at least as good as larger ones for learning what is common to a set of experiences. However, as we have already suggested, we remain unconvinced that such a distinction is necessary. It may turn out that learning of generalizable skills is simply insensitive to the size of the changes made in connection strengths.

[3] As noted previously, spared learning effects also show up in single-trial priming experiments. One view of these effects, consistent with the separate systems view, is that they reflect the subtle effects of single trials in those parts of the system where skills and procedures are learned. Again, an alternative would simply be that the priming task is less sensitive to the magnitude of the changes in connection strengths and more sensitive to the relative size of changes made by different stimuli.

CONCLUSION

In this chapter, we have considered the phenomenon of bitemporal amnesia in the light of models of distributed memory. We have described a hypothetical mechanism that can account for temporally graded retrograde amnesia without assuming that recent memories are stored separately from older ones. We have demonstrated how the ability to learn gradually from repeated experiences is an automatic consequence of assuming that amnesia simply amounts to the reduction of the effective size of the changes in the connections in a distributed memory. And we have indicated how a distributed approach can allow us to suggest reasons why large changes to connection strengths might make more of a difference in forming explicit representations of facts and episodes than in laying down the connections required for cognitive skills.

Obviously, there is considerable room for further work to test and to extend our views. If our hypotheses are correct, and if γ really is a chemical, then we might hope that someday someone may discover just what the chemical is, and will then go on to show that normal memory depends only on γ and not on some information processing activity that takes place in the hippocampus, as has frequently been suggested (Squire et al., 1984; Wickelgren, 1979). Considerably more theoretical work will be required to build a tight connection between those tasks in which spared learning is observed empirically and the situations in which large increments to the weights do not result in superior learning. In the meantime, we hope this chapter has demonstrated that what we know about amnesia is not only consistent with the idea of distributed, superpositional memory, but that certain aspects of the amnesic syndrome—in particular, residual learning in domains where amnesics show deficits—actually support the idea.

ACKNOWLEDGMENTS

The work reported here was supported in part by a grant from the System Development Foundation, in part by contracts from the Office of Naval Research (N00014-79-C-0323, NR667-437 and N00014-82-C-0374, NR 667-483), and in part by a NIMH Career Development Award (MH00385) to the first author. We would like to thank Neal Cohen, Morris Moscovitch, Daniel Schacter, and Larry Squire for several useful discussions of various aspects of amnesia on several different occasions.