Memory Distortion

How Minds, Brains, and Societies Reconstruct the Past

Editor
Daniel L. Schacter

Contributing Editors
Joseph T. Coyle
Gerald D. Fischbach
Marek-Marsel Mesulam
Lawrence E. Sullivan

Harvard University Press

Cambridge, Massachusetts
London, England
1995

Constructive Memory and Memory Distortions: A Parallel-Distributed Processing Approach

2

James L. McClelland

Bartlett (1932) introduced and insisted on the view that memory is a constructive process. His view was essentially that recall is not a retrieval, but a reconstruction, in which aspects of the content of previously presented material are woven into a coherent whole, with the aid of preexisting knowledge. Details may be distorted to increase coherence; rationalizations not present in the original may be introduced; details that are consistent with the synthesized coherent story may be added; and details that are inconsistent may be dropped. Neisser (1967) likened both perception and memory to the constructive activities of a paleontologist, who uses a collection of bone fragments, as well as everything she knows about dinosaurs from previous experience, to reconstruct the skeleton of a particular dinosaur. These ideas are consistent with what we would refer to today as a constraint satisfaction process, in which remembering is simultaneously constrained by traces left in the mind by the event we are remembering itself, by background knowledge of related material, and by constraints and influences imposed by the situation surrounding the act of recollection. Obviously if memory is constructive in this way, this has profound implications for the question of the veridicality of memory and the extent to which it may be influenced by suggestion, preexisting knowledge, and other related experiences.

My interest is in the mechanisms that may implement this constructive, constraint satisfaction process. Remembering, I will argue, takes place in a parallel distributed processing system—a system consisting of a large number of simple but massively interconnected processing units. Processing in such systems takes place through the propagation of activation among the units, based on excitatory and inhibitory connections. Forming a memory trace for something—say, an episode or event—begins with the construction

of a pattern of activity over the processing units, with the experience itself strongly influencing the pattern. But the existing connections among the units will also influence the pattern constructed, thereby introducing the possibility of additions, omissions, and distortions. Storage of a trace of the episode or event then occurs through the modification of the strengths of the connections among the units; to a first approximation, the connection from a unit that is active in the representation to another such active unit will tend to increase in strength, while the strength of connections between active and inactive units will tend to decrease.

Remembering may occur when some aspect or aspects of an event arise again as input. This may activate some of the units that previously participated in the representation of the episode or event, and these may in turn activate other units, via the weighted connections. The pattern that is constructed again depends on the connections among the units, and since these were adjusted previously when the episode was first experienced, the pattern that is constructed will tend to correspond to the pattern that was present at the time of storage. But the units that participate in representing one episode or event also participate in representing other episodes, and so the representation that is constructed may be affected by many other experiences. This means that my memory of any one episode or event will tend to reflect the influence of what I have learned from many other episodes or events.

I will describe two models that capture this constructive process in different ways. Both models have their origins in early connectionist papers, one by myself (McClelland, 1981) and one by Hinton (1981). Neither model is fully adequate in itself, but I will propose a synthesis of the two that may capture some of the main features of human memory, including aspects of memory distortions. The synthesis may provide as well one way of thinking about amnesia.

A Trace Synthesis Model

The first model (McClelland, 1981) focuses on distortion processes that can occur during acts of remembering, using a simple, localist connectionist network for the storage and retrieval of information. I used in the example the task of remembering facts about a collection of two somewhat unsavory individuals, belonging to two made-up gangs, the Jets and the Sharks. The Jets tended to be in their twenties, to be single, and to have only a Junior High School education, though no one Jet had all these characteristics; the Sharks tended to be older, to be married, and to have attended High School, though again no one Shark had all these properties. Members of both gangs were equally likely to be pushers, bookies, or burglars.

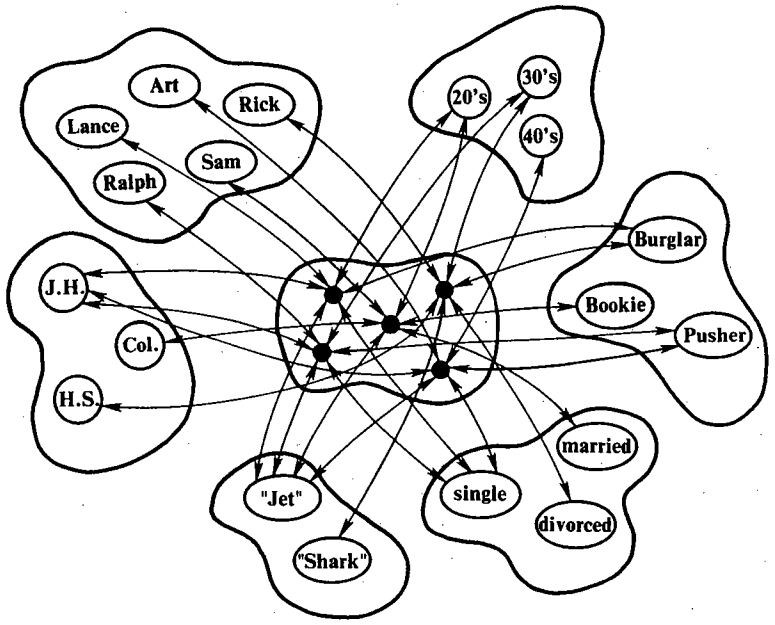

Figure 2.1 The network used to demonstrate aspects of constructive memory by McClelland (1981). The units participating in the representation of a few of the members of the Jets and Sharks gangs described in the text are shown. Units within the same group are mutually inhibitory; units connected with bidirectional arrows are mutually excitatory. From McClelland (1981), fig. 1, p. 171.

In the model, I represented each individual with its own connectionist processing unit that I will call an instance unit (see Figure 2.1). The model also contained property units, one for each property an individual might have. Names were treated as properties, so there were units for names, for gang membership, for education, for marital status, and for occupation. Bidirectional, excitatory connections between units were used to link instance units to the units representing their properties, so that if one activated a name unit, it activated the corresponding instance unit, and the instance unit activated the other properties of the instance. The instance units formed a group of units that were mutually inhibitory, so that if one was active it tended to suppress the others; similarly, the units for each type of property were grouped into clusters of mutually inhibitory units. The use of single units to represent whole items is a simplification—I will argue later that it

is more correct to use distributed representations, both for the whole and for the parts. However, even this simple model captures crucial aspects of the kinds of reconstructive processes that take place during remembering.

My focus in the original work reported in McClelland (1981) was on the process of constructing representations of material not explicitly stored in memory. One such case involved constructing a composite recollection of the typical Jet or Shark. In the model this could be done simply by activating the unit for Jet. This unit then sent activation to the instance units for all of the Jets, and these in turn sent activation to the units for each instance's properties. The inhibition among the instance units prevented any of these units from becoming too strongly activated, but they all contributed some activation to their property units. As a result the properties of the typical Jet became active (age in 20's, single, JH education); all of the occupations were partially activated.

This first example shows a desirable property of this sort of memory system—it can spontaneously generalize from examples. Another property—which may often be desirable but which can also be undesirable—is revealed when the model is used to try to retrieve the properties of a single individual by activating the unit for that individual's name. This individual's properties tend to be more strongly activated than the properties of any other individual, but one finds that as the activation process goes on, other, similar individuals become partially activated. This happens because as the properties of the target individual become active, they send activation to the instance units for other individuals, and these, in turn, tend to activate the units for their properties. This effect can be particularly potent—and can lead to strong distortions—when some piece of information about the target individual is missing. To show this, I first deleted the connection between the instance node for Lance and the property node for his occupation—he happened to be a burglar. Then I activated the name unit for Lance. This caused the instance node to become active, and the instance node then activated the property nodes for Jet, 20's, JH education, and married. Now, there happened to be several other Jets who had many of these properties, and they all happened to be burglars. As a result, the model filled in this occupation for Lance. In this instance the result happened to be correct, but the same thing would of course have happened whether Lance had been a burglar or not. Had he been a pusher or a bookie (or, for that matter, someone with an entirely innocuous occupation), the model would have filled in burglar anyway. In that case this would have been a clear example of a memory distortion: Lance would have been guilty by association.

The model illustrates two key points central to the issues raised in this volume. First, it provides an explicit though simple mechanism illustrating how memory distortions can arise from the workings of ordinary memory retrieval processes. These processes are often beneficial—they allow the formation of generalizations over similar instances and the filling in of missing properties based on the properties of other, similar individuals—but they can potentially be harmful in that the information filled in need not be correct.

Second, the model has the same property that human memory has, of often failing to separate information that arises from different sources. Suppose that a new instance node is formed from every experience (a proposal strikingly similar to the memory model of Hintzman, 1988), and suppose one has a number of similar experiences. Then when we try to recall one, pieces of other similar experiences will tend to intrude particularly in those aspects of the original for which the information is weak or missing. In the model, it is unfortunately not possible to inspect each memory trace individually; the information is not stored in the units themselves, but in their connections; like connections among neurons in the brain, we only know what is stored in the connections through the effects these connections have on the outcome of processing. But, since many units and connections contribute to this outcome, full disentangling of the specific cause of each aspect of the outcome is impossible. It will, then, in general not be possible to identify the specific source of any aspect of constructed recollection.

Given the model, then, the memory distortions reported in this volume by Loftus, Ceci, Moscovitch, and others are to be expected. Perhaps the only thing that is unexpected about them is the resistance that often arises to their acceptance. This resistance may come from implicit acceptance of an alternative model of memory, in which memory traces are not so much constructed as retrieved, like books from a library. The metaphor of human memory and human knowledge as a library provides a basis for accounting for the role of organization in memory, but gives no basis for understanding distortion. I believe that as we come to understand memory better and better, it will become increasingly clear that this is a misleading metaphor.

An Experimental Test of the Trace Synthesis Model

To test the model described above, and to extend the empirical data base of evidence of memory distortions, Leigh Nystrom and I developed an experimental paradigm designed to elicit trace synthesis errors in remembering (Nystrom and McClelland, 1992). In this paradigm, subjects study a list of sentences, and then are later cued for complete recall of individual whole sentences when a fragment is presented as a probe. Consider sentences of the form: "The policeman gave the accountant the hammer in the basement." We imagine that the sentence is analyzed into a set of role-filler pairs, which are then represented by the activations of input units in the network shown in Figure 2.2. The network is strictly analogous to the model previously described. There is a pool of instance units, with one unit assigned to

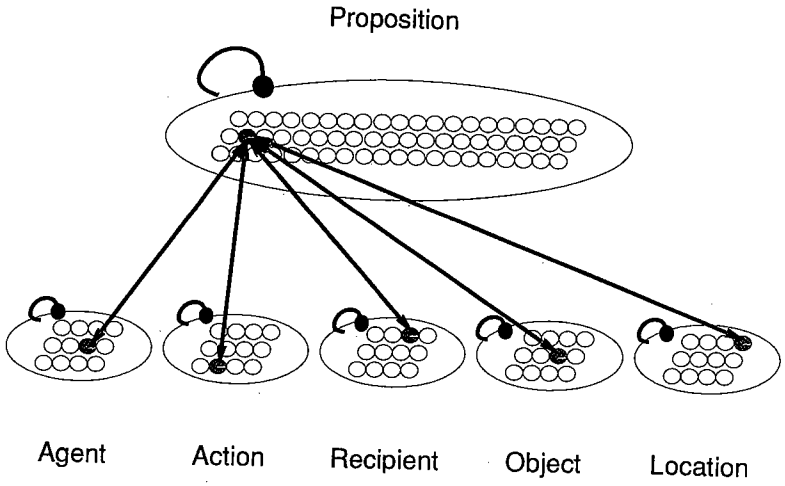

Figure 2.2 Sketch of the network architecture used by Nystrom and McClelland (1992). The units participating in the representation of a particular proposition are shown, along with the bidirectional connections that allow this model to perform cued recall of the whole sentence when part of the sentence is represented as a cue.

each proposition. The model also contains several pools of property units, one for each type of role that can occur in one of the sentences. Each pool contains a unit for each word that can appear in that role. We represent the sentence about the policeman as shown in the figure. Another sentence, with a non-overlapping set of words, would involve a different word-in-role unit in each property unit pool, and a different proposition unit.

In a model such as this, cued remembering occurs by simply turning on the units for the words contained in the probe, and asking the network to essentially fill in the rest. This occurs via a gradual constraint satisfaction process. Processing begins with the units for the cue words clamped and continues until a stable pattern of activation is achieved. When the cue uniquely matches one stored sentence, and there are no other very similar sentences, the correct sentence tends to be recalled. However, errors can occur when there are two or more stored items that have the same or a similar degree of match to the probe. In this case the "remembered" pattern is a constructed synthesis of two or more stored traces.

Nystrom and I studied the adequacy of this model to account for memory and memory distortions in a series of four experiments. Here I will discuss only one of these. The subjects studied lists of sentences set up so that some of the sentences shared three content words in common with another paired

Table 2.1 Example overlap and control sentence pairs from Nystrom and McClelland (1992), with corresponding test probes

Overlap Pair:
The policeman gave the accountant the hammer in the dining room.
The farmer gave the accountant the hammer in the garage.
Control Pair:
The driver showed the receptionist the toaster in the kitchen.
The swimmer loaned the salesman the envelope in the basement.
Overlap Probe:
The . . . gave the accountant the hammer in the . . .
Control Probe:
The . . . showed the receptionist the toaster in the . . .

sentence. One such pair of overlapping sentences is shown in Table 2.1; a pair of control sentences, with no overlap, is also shown. In the model, we assigned a different proposition unit to each sentence, and connected this unit to the input units for each of the corresponding words. In the case of overlap sentences, three of the five role-fillers are the same, so the proposition units in these cases are connected to overlapping sets of input units, as shown in Figure 2.3.

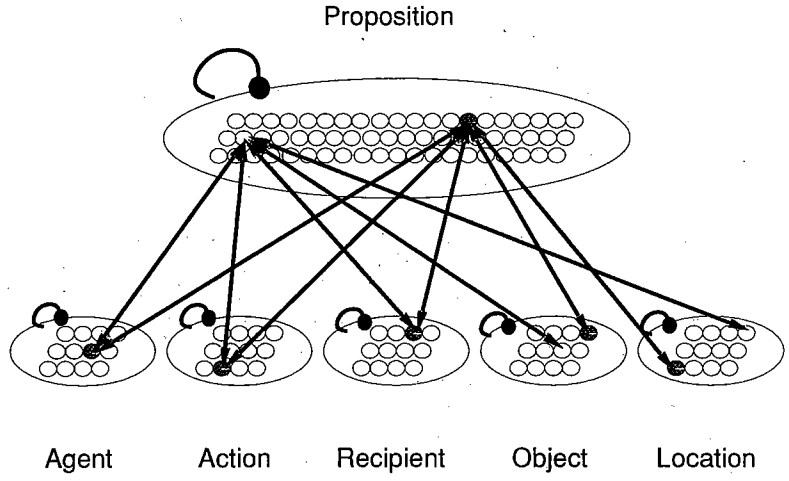

Figure 2.3 The units and connections participating in the representation of a second sentence that overlaps with the first one shown previously in Figure 2.2. Presentation of an ambiguous probe tends to activate both sentences and may produce a blend error.

Our focus in this research was on trace synthesis at the time of recall, and we therefore went to some lengths to minimize the possibility that subjects would be reminded of the first member of each overlap pair when given the second member during the exposure phase of the experiment. This was done, first of all, by developing a cover task that focused subjects on analyzing each individual sentence separately from all of the others without any mention of a later memory test: subjects were told their task was to rate each sentence for its overall plausibility and to say how well they thought each word fit with the overall event described by the whole sentence. Other precautions included varying the placement within the sentences of the overlapping words, and separating overlapping sentences as far as possible in the study list of 34 sentences. Although subjects did notice that some words were occasionally repeated, only a few subjects reported that the second member of an overlap pair ever caused them to recall the previous member of the pair, and eliminating these subjects from the analyses did not change the results. Thus we were reasonably confident that reminding and trace synthesis during the study phase was not a major factor in determining the results.

After the subjects completed a filler task for 5 minutes, the cued recall phase of the experiment was administered. This involved presenting 16 sentence fragments, each with blanks for two content words. Eight fragments were from overlap pairs and eight were from control pairs. Thus a subject might see probes like the ones shown in Table 2.1.

In both cases, the subjects were told to complete the probe with the first studied sentence that came to mind. Subjects were alerted to the fact that sometimes the probe matched two studied sentences equally well, and they were told to recall only one of the two sentences and to take care not to mix up the two. After the first recall they were given an opportunity to recall the second sentence. I will be discussing only the results of the first recall here. Suffice it to say that second recalls were generally less accurate than first recalls.

On the first recalls, the probability of correctly recalling a complete sentence did not differ between the overlap and control pairs: two words from the same sentence that matched the probe were recalled 42% of the time in the overlap condition and 41% of the time in the control condition. However, in the overlap condition subjects did sometimes make what we can call synthesis errors—errors in which one word came from one of the sentences that matched the probe and the other came from the other of these two sentences. This occurred on nearly 10% of the error trials (5.4% of all trials with overlap probes, compared to less than 1% of trials with control probes). The rate of synthesis errors may seem relatively low, but they were reliably more frequent than chance. Confidence ratings were obtained on each recall trial; confidence was slightly less on average for synthesis errors

than for completely correct responses, but on 40% of the synthesis errors subjects gave the highest confidence rating, corresponding to the statement "I am sure both words recalled came from a single studied sentence that matched the probe." We take the experiment, then, as demonstrating that memory distortions can arise from trace synthesis at the time of recall. We would not want to claim, of course, that trace synthesis does not often occur earlier, when an intervening event reminds us of a previous event; indeed, it may be that this is one common source of memory distortions. We would only suggest that our results support the view that trace synthesis can occur at recall as well as between initial study and recall.

We modeled the data from this and the other three studies we conducted using the model discussed previously. To fit the data it was necessary to make two additional stipulations: first, that processing has an inherently random component; and second, that subjects sometimes failed to encode each sentence completely. We implemented this latter assumption by randomly eliminating a fraction of the connections between the input and prop units. These assumptions do different and important work in accounting for the data.

The first assumption—intrinsic variability—allows the network to select essentially randomly between two equally good responses in cases where two studied sentences fit the probe equally well. Intrinsic variability is implemented simply by incorporating normally distributed random noise into the input to each unit. Each time the unit's state is updated, this noise affects the exact degree of activation. If high levels of noise are used, the model becomes totally random in its behavior; but with small amounts of noise, the variability effectively causes the network to choose randomly among equally good alternatives. Without any noise the network will have a tendency to partially activate both matching sentences most of the time, and will not tend to recall one or the other: with the randomness in place, on the other hand, the network will tend to settle to one of the two answers. The idea of intrinsic variability in processing was introduced into connectionist modeling by several investigators in the mid-1980s (Geman and Geman, 1984; Smolensky, 1986; Hinton and Sejnowski, 1986). I have suggested elsewhere that intrinsic variability is a general property of human cognitive function, and I think it is necessary if we are to model the kinds of results we see in a wide range of tasks, such as free association, stem completion, or perceptual identification, where subjects generally emit one or the other of a set of alternative coherent responses, rather than a blend of many alternatives (McClelland, 1991). Others have established that the outcome of this settling process is optimal from the point of view of maximizing the probability of selecting the correct answer, given that the weights accurately encode information about the domain (Geman and Geman, 1984).

The second assumption—encoding failures—allows the model to ac-

count simultaneously for the existence of synthesis errors, together with the fact that the probability of correct recall did not suffer in the overlap condition compared to the control condition. Simplifying a bit, with this assumption in place, correct recall of a single sentence depends on whether it has been completely encoded, and the probability of complete encoding is independent of whether there is an overlap sentence in the study set. Incomplete encoding offers an opportunity for synthesis errors: we obtained an excellent fit to the data by assuming that subjects failed to encode 20% of the words. In cases where there are gaps in the encoding of one of the sentences, the other can contribute, creating a memory distortion. Intrusions from the other sentence rush in when the most active trace provides no information.

Summary of the Trace Synthesis Model

The model I have described has considerable appeal as a simple descriptive account of the process of memory trace synthesis in cued recall and goes some way toward implementing the constructive memory retrieval process of which Bartlett and Neisser wrote. I should note that there are other models that can account for trace synthesis, including the MINERVA model of Hintzman (1988), as well as the models by Metcalfe (1990) and by Humphreys, Bain, and Pike (1989) that use distributed representations (see also McClelland and Rumelhart, 1985). The model of Metcalfe (1990) has been applied to a number of important findings on blend errors and other memory distortions and shows that such models can offer very nice accounts of much of the existing data on blending and memory distortion.

In spite of their success, all of these models lack something. There are other, deeper, more fundamental forces at work shaping memory performance. These processes, I believe, operate gradually over the course of cognitive development to shape the way we represent the constituents of memory traces—for example, the concepts that contribute to propositions. These representations, in turn, provide the basis for more powerful forms of constructive memory effects.

Models of Representation Formation via Gradual Learning

An early model that pointed the way toward this idea was presented by Hinton (1981). This model is sketched in Figure 2.4. It consists of separate sets of units for representing the first noun, relation, and second noun of three-term propositions such as "Fish can swim," "Sammy is a fish," "Elephants are gray," "Clyde is white," and so forth. The model is similar to the previous model, but now each word is a pattern of activation over the appropriate units rather than a single active unit. The whole proposition is represented as a pattern of activity over the three sets of constituent units,

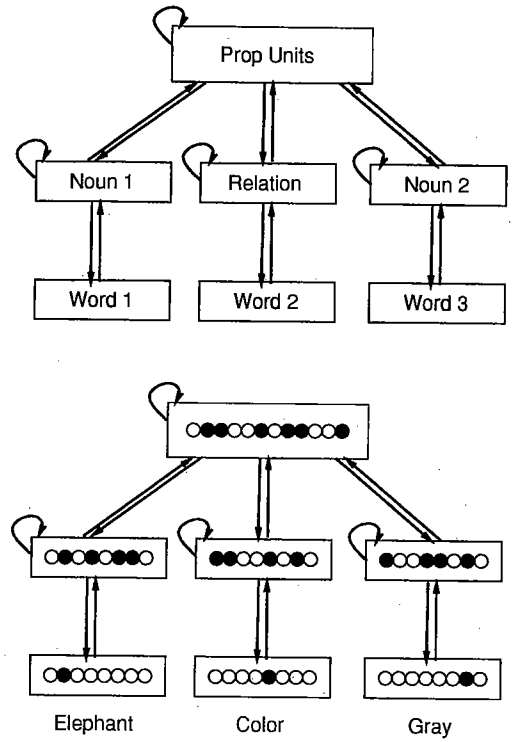

Figure 2.4 A slightly elaborated version of the connectionist network used by Hinton to represent propositions about the concepts. Each rectangle represents a pool of connectionist information processing units, and each arrow represents a full set of connections from each unit on the sending side of the arrow to each unit on the receiving side. Processing occurs by the propagation of activation among the units via the connection weights. Each unit simply sets its activation to a value between 0 and 1 based on the summed input it receives from other units via the weighted connections. For example, the connections from the Word 1 input units to the Noun 1 representation units allow a pattern of activation representing the first word of a proposition to activate the appropriate semantic pattern over the Noun 1 units. The connections between the different pools of constituent units and the Prop units encode the system's knowledge about propositions, and after the weights have been acquired through learning, these connections allow the third constituent of a proposition to be completed given the other two constituents as inputs. Return connections from the constituent units to the word units then provide for output of the pattern filled in by the network. The upper panel shows each pool of units with labels. The lower panel illustrates how a particular proposition would actually be represented.

and over an additional, fourth set of units called "PROP" units. The network contains bidirectional connections from each set of constituent units to the PROP units. There is also another set of connections for input to and output from the network; these allow inputs standing for specific words to activate distributed semantic patterns over the input units. The recurrent connections within each pool of units allow local pattern completion within each pool. The effect of this is to implement a "clean-up" process in which the pattern of activation tends to converge to the representation of a specific word, and has much the same effect as the mutual inhibition within layers in the previous model.

Once again, the knowledge or memory in this system is stored in the connections among the units. We can think of the input/output weights as encoding knowledge about the semantic pattern corresponding to each word of the proposition, and we can think of the connection weights between the constituent and PROP pools as encoding knowledge about the propositions that these constituents enter into with other constituents. Hinton (1981) suggested that this network would be able to generalize what it knows about one concept to other related concepts if similar concepts are represented by similar patterns of activation. Thus, if Clyde is a particular elephant, and Clyde is represented by a pattern that is similar to the pattern for elephant, then what we know about elephants will tend to generalize to Clyde. Such effects do not strictly obey the laws of logic; instead they obey the laws of association.

In Hinton's (1981) work, the representations of the concepts were assigned by hand. Connectionist learning algorithms have evolved considerably since that time, however, and we now have algorithms that can discover how to represent different concepts through repeated exposure to information about the entire semantic domain in which the concept is embedded. I will consider one such domain—the broad domain of living things. I show in Figure 2.5 a representation of a fragment of the knowledge someone might have about living things. This format is typical of the approach to knowledge representation used in classical artificial intelligence approaches to cognition, beginning with Quillian (1968). The knowledge has several characteristics: it is structured, in that it is organized into a hierarchy. Individual types or species are listed at the bottom of the hierarchy, and their organization into broad classes, and the organizations of these into larger classes, is indicated by "isa" links. We can imagine attaching, below the level of the types, specific instances of the types. For example, we could add "Tweety isa canary," and so forth; or if we had an Elephant node, we could add "Clyde isa Elephant." The network is potentially quite economical, in that facts that are true of whole sub-trees of the hierarchy can be attached at the highest level to which they apply. Given this, when some information is not stored on a specific concept, it may be inferred by searching through the tree. The process is equivalent to the standard logical syllogistic reasoning process through which we infer that Socrates is mortal. We know that Socrates is a man, and we know that men are mortal, therefore we can infer that Socrates is mortal too.

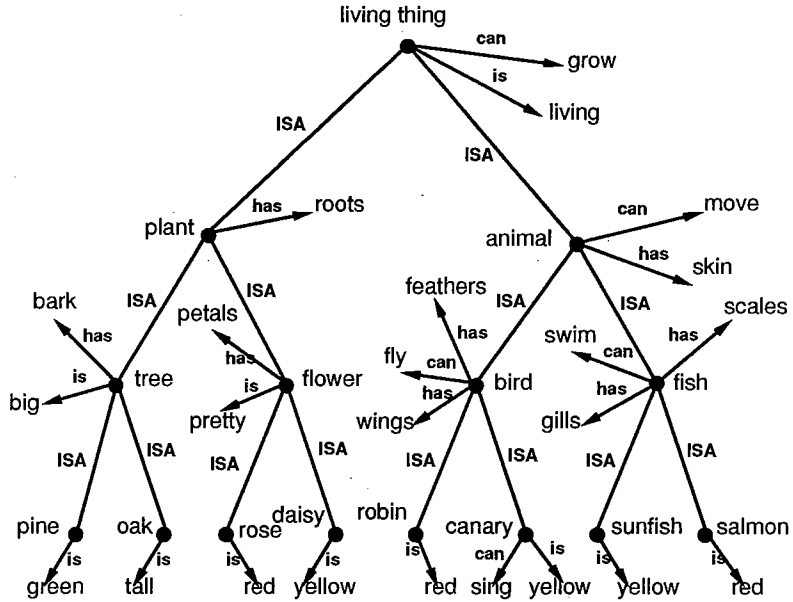

Figure 2.5 A semantic network of the type formerly used in models of the organization of knowledge in memory, containing the concepts and propositions used in the learning experiment of Rumelhart (1990). Adapted from fig. 1.8 of Rumelhart and Todd (1993) by McClelland et al. (1994) as their fig. 3, p. 13.

When one trains a network like the one shown in the previous figure with example propositions from this domain, it learns two things. It learns connection weights internally that encode the propositions, and that allow completion of a proposition from two of its terms. It also learns connection weights from the word input units to the constituent units that essentially assign useful semantic representations to each word. Hinton (1989) demonstrated this for kinship relationships. Rumelhart (1990; Rumelhart and Todd, 1993) applied the same idea to the domain of living things (the actual simulation model Rumelhart used was slightly simpler than the one shown in Figure 2.5), and I have chosen to use this case as my example. The results on which the following discussion depends come from a repetition of the Rumelhart (1990) simulation reported in McClelland, McNaughton, and O'Reilly (1994).

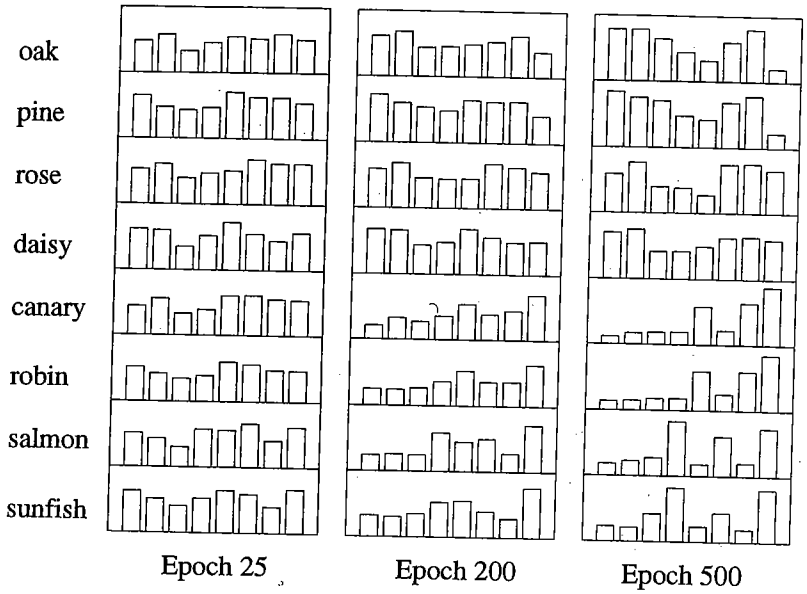

Figure 2.6 Representations discovered in a replication of Rumelhart's (1990) experiment training a semantic network much like the one shown in Figure 2.5. The figure shows the activation of each of the Noun-1 units for each of the eight specific concepts used. The height of each vertical bar indicates the activation of the unit on a scale from 0 to 1. One can see that initially all the concepts have fairly similar representations. After 200 epochs of training, there is a clear differentiation of the representations of the plants and animals. After 500 epochs, the further differentiation of the plants into trees and flowers and of the animals into fish and birds is apparent. From McClelland et al. (1994), fig. 5, p. 16.

Through gradual training on examples from the domain of plants and animals, the network learned more than just the propositions. It also learned to assign useful representations to each concept. The representations the network learned to use for the first noun are illustrated in Figures 2.6 and 2.7. These representations are determined by the connection weights from the concept input units to the concept representation units. Through the course of learning, these weights gradually change, so that the representations of the different concepts gradually come to capture how similar the concepts are in terms of the propositions they enter into. Canary and Robin enter into highly overlapping sets of propositions—for example, both are birds, both can fly, both have feathers. As a result of this, the network comes to assign them representations that are very similar; similar representations lead to similar outputs. Most important, once it has learned to use such

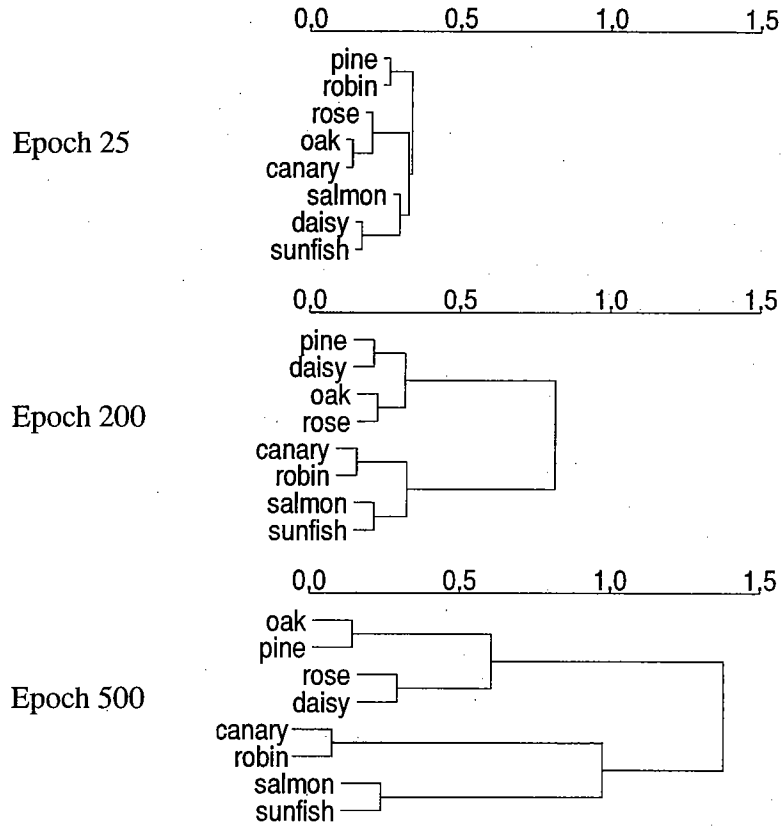

Figure 2.7 Similarity structure discovered in our replication of Rumelhart's (1990) learning experiment, using the representations shown in Figure 2.6. Initially, the patterns are all quite similar, and the weak similarity structure that exists is random. The concepts become progressively differentiated as learning progresses. From McClelland et al. (1994), fig. 6, p. 17.

representations, the network can use the similarity of the representations as the basis for inferences. Thus, once the network has learned how to represent Canary and Robin as similar to each other and distinct from other concepts, it now adds to the training set one proposition about a new type of animal—let's say the proposition that "Sparrow isa bird"—the network learns to assign "Sparrow" a representation similar to the representation it uses for other birds. After this proposition has been learned, we can then ask the

network if it knows what a sparrow can do. This can be done by pattern completion—we can test the network to see if it can complete the pattern "Sparrow can . . ." with "fly." Indeed, Rumelhart showed in his simulations that if the network was trained on the full set of propositions concerning canaries and robins, he could teach it only one proposition about sparrows—namely "Sparrow is a bird"—and it was able to correctly complete other propositions about sparrows. The output was quite clear about those things that are generally true of the other birds. For those properties that differed between canary and robin, it produced ambiguous outputs. Thus it applied what it already knows about canaries and robins to sparrows.

Now, taking this model at least as a sketch of a model of our knowledge of facts and events, let us consider the nature of memory as reconstruction again. Individual experiences themselves are not separately represented; instead, they leave what I would call a structured system of knowledge stored in the connection weights. Furthermore, this knowledge is not itself directly accessible to overt responding or direct report. Instead the knowledge provides a mechanism that can construct responses to queries presented to the network, whether the actual proposition was ever actually experienced, as in the case of the actual training examples, or not, as in the case of questions we may ask about, for example, what a sparrow can do after the training described above. The outputs of such a network might then be the basis of performance we take as indicative of remembering, but many times they might reflect generalization based on the accumulated effects of prior experience, rather than the effects of storing anything like the specific item in memory. Such generalization is, I would suggest, central to our ability to act intelligently, and the process of learning the sorts of representations on which such generalizations are based is central to cognitive development (see McClelland, 1994, for discussion). But such generalization gives rise to distortions as an inherent by-product: it becomes impossible to distinguish between what has actually been experienced and what can be constructed based on other related things that have been experienced.

But there is something slightly wrong here. Our ability to isolate particular memories is not as bad as it would appear to be if we assumed that memory consists solely of the gradually acquired residue of a large body of experience. I can tell you a new fact—such as the fact that "Sammy is a Sunfish"—and this can affect your semantic memory right away. We need some mechanism capable of relatively rapid learning of the contents of individual episodes and experiences.

One might think that one could simply add new memories one at a time into a network like the Hinton (1981) network, but in fact this is not so. If one attempts to store additional memories all at once in such systems, it can be done, but at the cost of a phenomenon called "Catastrophic Interference" (McCloskey and Cohen, 1989). The addition of the new material

causes a dramatic loss of the ability to perform correctly with other, similar material, particularly when, as is often the case, the new material is not completely consistent with what is already known. Thus if I train the network with the propositions "Sparrow isa bird" and "Sparrow is brown," it will drastically interfere with my ability to recover the color of other birds like canaries and robins (McClelland et al., 1994). The only way to add new information robustly to a structured memory system is to add it through a process called interleaved learning, in which learning occurs very gradually through repeated exposure to the new material, interleaved with ongoing exposure to other examples of the same domain of knowledge. Connection weight adjustments occur during exposure to the new material and the old, thereby gradually allowing the new material to be incorporated into the memory system without at the same time disrupting what is already known.

A Proposed Synthesis of the Models

A natural proposal that arises from this observation, then, is to suggest that the human memory is essentially a synthesis of the two types of models I have described above. One part of the system gradually learns to represent and use concepts as in the Hinton model, while another part is given the task of rapidly learning the specific content of individual events and experiences, storing them in a way that is similar to the method used in the Trace Synthesis model of McClelland (1981). I have presented a visualization of this idea in Figure 2.8.

This proposal may seem at first somewhat unparsimonious, but in fact it provides an account of the pattern of amnesia that results from bilateral lesions to the medial temporal lobes. Individuals with extensive damage to these brain regions show a very striking pattern of memory deficits (for overviews see Squire, Chapter 7 of this volume). These patients appear profoundly deficient in the ability to form new semantic or episodic memories, but the ability to acquire new implicit knowledge such as new cognitive skills or sensitivity to the sequential dependencies among stimuli in implicit learning tasks remains intact, and existing semantic knowledge such as semantic associations can be primed. Semantic and episodic knowledge acquired long before the damage occurred is spared—that is, it is as good in such patients as it is in age-matched controls. In fact, there is a temporally graded retrograde amnesia, which in humans can extend over several years, such that semantic and episodic memories that were acquired shortly before the occurrence of the damage are profoundly affected, and memories that were acquired at progressively earlier times are progressively less and less affected. Crucially, in several studies both in humans and in other animals, memory for the most recent premorbid time periods can actually be much worse than memory for material from slightly more remote time periods.

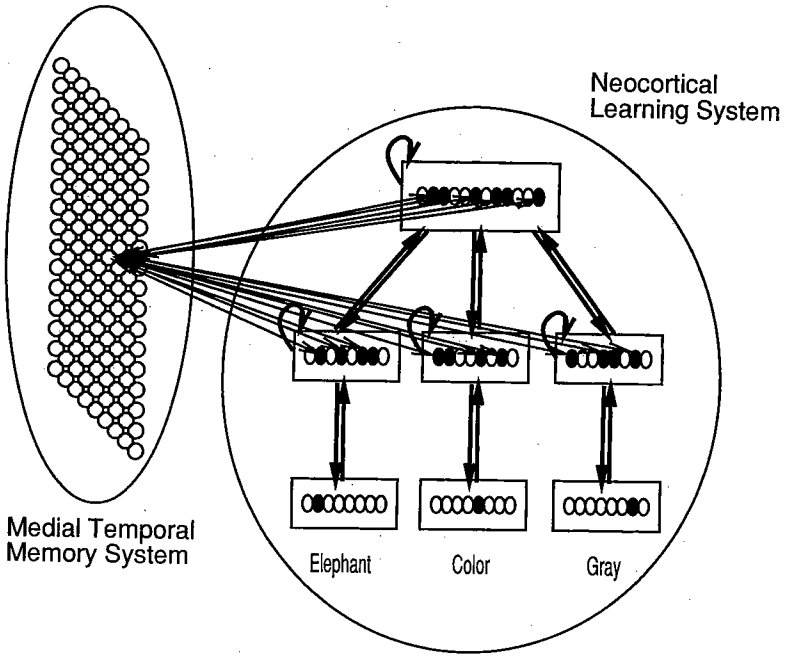

Figure 2.8 A synthesis of the McClelland (1981) model and the Hinton (1981) model. One part, based on the McClelland (1981) model, allows the storage and retrieval of individual traces, subject to trace synthesis, while the other part, more similar to the model of Hinton (1981), makes use of gradual, interleaved learning to acquire a structured system of knowledge gradually from exposures to ensembles of events and experiences. The part of the model based on McClelland (1981) plays a role akin to that played by medial temporal lobe structures in the brain, while the part based on Hinton (1981) plays a role similar to that played by other learning systems in the human neocortex.

We can account for these findings by assuming that older, consolidated memories, as well as cognitive skills and other "implicit" forms of knowledge, are subserved by information processing systems located in the large neocortical information processing system situated outside the medial temporal region. I identify these systems—hereafter labeled collectively the neocortical system—with the systems that acquire knowledge very gradually, through small adjustments to connection weights, represented in Figure 2.8 schematically by the network of the type introduced by Hinton (1981). The connection adjustments in this system, as we have seen, lead to the gradual emergence of structured knowledge systems such as those that are required for adequate generalization in domains that others have tended to treat as implicit—domains such as syntax—and domains that others have tended

to treat as explicit—domains such as semantic memory. At the same time we assume that the ability to perform correctly in explicit memory tasks based on rapidly formed memory traces of recent events and experiences arises from learning that takes place within the medial temporal lobes, hereafter called the medial temporal lobe system. Recently, Bruce McNaughton, Randy O'Reilly, and I have proposed an account of the amnesic syndrome based on these ideas (McClelland et al., 1994). On this view, an experience, such as hearing someone say "Sammy is a sunfish," produces a pattern of activation widely distributed throughout the neocortical system; connections from this system into the medial temporal region produce a corresponding pattern of activation over the neurons there. The medial temporal lobe system then plays the role that the proposition units play in the McClelland (1981) model, linking all of the constituents of the event together into a single trace. We do not think this is done by assigning an individual neuron to each episodic memory, as originally proposed by McClelland (1981). However, an explication of the details of our view of the nature of medial temporal lobe representation is beyond the scope of this chapter. Suffice it to say that we think of the representation as sharing many characteristics with the representations used in the trace synthesis model: the representations, though distributed, are relatively sparse (few units active), and each unit that participates in the representation is activated only when a conjunction of elements occurs in the input (for fuller discussion, see O'Reilly and McClelland, in press).

Once a representation has been set up in the medial temporal lobe system, memory can be probed by presenting an incomplete fragment, just as in the trace synthesis model, and reconstruction occurs, via return connections. Each time a trace is synthesized, a small amount of connection adjustment takes place within the neocortical system as well. Consolidation is thought to be the result of this gradual neocortical learning that occurs every time a memory trace is reconstructed. The process is gradual, so that the new information initially stored via the medial temporal system itself can be gradually integrated into the system of representations used in the neocortical system without disrupting existing knowledge stored therein. This sort of dual memory system then allows new information to be rapidly stored in the medial temporal system without producing catastrophic interference with what is known in the neocortex. Information that is repeatedly reinstated, interleaved with ongoing exposure to other information, gradually becomes incorporated into the representations in the neocortex.

Conclusion

I have concurred with those who hold that memory is a constructive process, and I have proposed two rather different types of connectionist mechanisms that can contribute to the synthesis of memory traces; and I have suggested