MEMORY DYSFUNCTIONS: AN INTEGRATION OF ANIMAL AND HUMAN RESEARCH FROM PRECLINICAL AND CLINICAL PERSPECTIVES Vol. 444
Reprinted from
ANNALS OF THE NEW YORK ACADEMY OF SCIENCES

Distributed Models of Cognitive Processes

Applications to Learning and Memory[a]

JAMES L. McCLELLAND

Department of Psychology
Carnegie-Mellon University
Pittsburgh, Pennsylvania 15213

Several researchers, coming from a number of different backgrounds and viewpoints, have recently begun to think in terms of models of information processing and memory in which processing takes place through the simultaneous interaction of a very large number of very simple, but highly interconnected processing elements. The present paper gives a general characterization of this approach, describes some of the reasons for its appeal, considers how it relates to underlying physiological mechanisms, and discusses the implications of this view for our understanding of amnesia.

First, I want to make it clear that this approach is based on the work of many people. Perhaps most important has been the work described in the volume edited by Hinton and Anderson entitled *Parallel Models of Associative Memory.*[1] My own work on these models has been the result of a collaboration with David E. Rumelhart; most of the fruit of that collaboration will appear in a forthcoming book, though some of it is summarized here. It is also important to bear in mind that the class of models I will be describing represents a family resemblance structure. They have no necessary or sufficient conditions, though they generally agree on a number of basic properties. Third, the development of models of this class, which I will simply call distributed models, is still in its infancy. The application of these ideas to the psychology of normal memory and amnesia is just getting under way.

BASIC CHARACTERISTICS OF DISTRIBUTED MODELS

Distributed models are abstract characterizations of processing systems. They are analogous to computer programs, in that they do not specify the detailed physical processes underlying information processing; rather they describe processing at a more abstract, computational level. They differ from computer programs, though, in three fundamental ways. (1) Processing is thought to emerge from the interactions of a large number of simple (abstract) computational elements called nodes, rather than through the activity of a single, central processing unit. (2) The knowledge that underlies processing in distributed models is thought to be stored in the strengths of the connections between the nodes, rather than in data structures or compiled computer code interpreted by the central processing unit. (3) Learning in distributed models amounts to changes in the strengths of connections between the nodes, rather than changing the contents of data structures or recompilation of the program.

[a] Supported by a grant from the Systems Development Foundation. The author is a recipient of a Research Scientist Career Development Award (MH 00385).

What Are the Models Models Of?

A major issue that we must be clear about is what these models are models of. A major problem with answering this question is that different people answer it in different ways. My own view is that distributed models provide nice descriptions of the microstructure both of information processing and of the physiological substrate that underlies it. However, they do not describe the macrostructure at either level. At the cognitive level, this means that they do not describe the overall organization of behavior, though they are assumed to describe the structure of the primitive operations out of which the overall structure of behavior is built. At the physiological level, they do not describe the architecture of the brain in terms of the functional roles of different regions of the brain.

The fact that distributed models can be used to provide descriptions at two different levels of description does not mean that one and the same model will provide a satisfactory description at both levels, but simply that members of the same general class of models can characterize both levels. The exact relation between the two levels is, of course, a subject about which one could only speculate at this time. The models also have implications for how we should think about the macrostructure at either level. However, the laws that characterize the macrostructure of human performance may best be cast at a higher level of description. This is analogous to the relation between the subatomic and the molecular levels in physics and chemistry. Presumably, all the properties of molecular behavior have their basis in subatomic processes, but there are lawful regularities of molecular behavior that are conveniently captured, without resorting to a description of subatomic processes. Similarly, a description of the functions of the different regions of the brain need not refer in detail to the fact that each region is composed of neurons and synapses; and a description of the functions of each member of a sequence of information-processing steps need not refer to the fact that each step is based on the interactions of a large number of simple processing units all operating concurrently.

The fact that distributed models may be used to capture two different levels of description requires us to be clear about which level we are describing at any given time. In the first part of what follows, I will be describing models cast at the more abstract level, though when we turn to a consideration of amnesia, we will examine the relevance of the assumption that the same general principles apply at the physiological, as well as the psychological levels.

The remainder of this paper considers the relevance of distributed models to psychological and physiological theories of memory. The first section relates the approach to basic concepts in the psychology of memory. The second section mentions several reasons why the approach may prove to be appealing to psychologists. The third discusses the relation between distributed models at the psychological level and the possible physiological substrate of memory. The fourth section considers the phenomena of amnesia from the point of view of the distributed approach.

BASIC CONCEPTS IN MEMORY

The study of the psychology of memory has lead to a number of basic concepts. These have arisen from a variety of different theoretical contexts. Some of these concepts map easily onto the distributed framework, others do not. Here space only allows us to consider a number of concepts that can be easily related to this framework.

Current Representation or Mental State as Pattern of Activation

In distributed models, we think of a mental state as a pattern of activation over the set of computational elements whose activity underlies processing. This state is evolving and changing all the time, of course.

Let us note that the pattern of activation at any one time will reflect both the content of particular recent inputs to the memory system, for example a sentence an experimenter has given a subject to memorize, and the context in which that input occurs. For conceptual clarity, it helps to assume that the units in the model can be subdivided into those whose state is directly influenced by the content of the particular stimulus and those influenced primarily by the context; with the proviso that these units are interconnected with each other, so that the content conditions the context and vice versa.

It is also important to note that the nodes in a distributed model play particular roles in the patterns of activation. For example, in a system for processing words, some of the nodes will stand for different aspects of the sounds of the words and some for different aspects of their meanings. Different sounds are represented by alternative patterns of activation over the same set of sound nodes and different meanings by alternative patterns of activation over the same set of meaning nodes. It may be simpler to understand the models if we simply think of each node as standing for a particular attribute the item represented may or may not have; if the node is active, the item currently represented has the attribute, or at least the representation stipulates that it does.

Processing as an Interactive Process

Processing in distributed models results from the interact ons of the processing elements themselves. The result of the interactive activation process is the evolution of the patterns of activation through time.

The pattern of activation produced by any given input will be widely distributed over a large number of different parts of the overall cognitive system; for example, if the input is a written sentence, it will be distributed over primary and secondary visual areas, language areas, and areas specifically involved in mediating the connections between the two. Some of these processing areas can be thought of as playing a role in producing the higher levels of the representation; others can be thought of as being the ones in which those higher level representations are formed. However, viewed in another way, all the nodes in all of the different parts of the system play a role, both in the representation and in the processing that forms and maintains that representation. Fundamentally, the processing elements are the representational elements and the representational elements are the processing elements.

The Long-Term Memory Trace

Patterns of activation come and go, and are replaced by subsequent patterns of activation. What these patterns leave behind is not a copy of the pattern of activation itself, but changes in the strengths of interconnections among the basic processing elements. These changes have the property that they associate the different parts of the patterns of activation; that is, the set of changes to the connection strengths resulting from a particular mental state is such that the changes tend to permit different parts of the state, when they recur, to reinstate other parts of it.

Two essential properties of the long-term memory traces are that they are distributed and superimposed on each other. By distributed, we simply mean that the connection strengths that underlie a particular pattern are distributed widely in the connections between all of the units that represent the pattern when it is active. By superimposed, we simply mean that the traces of many different patterns are superimposed in the same set of connection strengths.

Learning as Incrementation of the Connection Strengths

Learning, then, amounts to making changes in the strengths of interconnections. Such changes are assumed to underlie all types of learning, as we shall see when we discuss amnesia below.

Retrieval as Reinstatement

Retrieval of a memory trace amounts to reinstatement of a prior pattern of activation, using part of the pattern as a cue. Note that the cue can be part of the pattern itself (the content), or of the context, or a mixture of the two.

Recognition as Reinstatement of Context

By recognition, I refer here to recognition of previous occurrence in a particular context. We assume that this kind of a recognition involves an act of reinstatement too—in this case, the subject must reinstate the context of an item from the item, rather than the item from the context.

Procedural and Declarative Knowledge

It is quite clear that some of what we know we can talk about and reason explicitly about, and some of what we know we cannot. Our view is that these different aspects of our knowledge are both encoded in the cognitive systems as changes in the connections between units. The distinction is simply this: some of these units serve the special purpose of providing the substrate for recreating representations accessible to report and reasoning processes, while others are more directly imbedded in the internal structure of the cognitive system. Changes in interconnections in the latter parts of the cognitive system would influence the details of stimulus processing, such as the effectiveness of formation of a representation of a stimulus from a fragmentary input, but would not give rise to changes in overt reports.

It should be emphasized that some of the relations between familiar concepts and aspects of distributed models are more clear-cut than others. However, the ideas outlined in this section should begin to suggest a link between the constructs typically used in cognitive psychology and constructs intrinsic to the theoretical framework of distributed models of memory.

REASONS FOR THE APPEAL OF DISTRIBUTED MODELS

One reason why distributed activation models are appealing is that their relation to physiology is much easier to visualize than it is for other information-processing models. However, their appeal stems not only from their "physiological realizability" but also from the fact that they provide very attractive models at the psychological level. The most important feature of distributed activation models is that they generalize spontaneously. Consider what happens when a number of different exemplars of the same concept are presented for processing, in conjunction with the appropriate label. For example, consider what happens when a young child sees different bagels, each of which he sees in the context of some adult saying "bagel" (obviously this oversimplifies the learning situation, but let us assume that the same principles would apply in a more realistic case). Let's suppose that the sight of the bagel gives rise to a pattern of activation over one set of units, and the sound of the word gives rise to a pattern of activation over another set of units. After each learning experience of the sight of a bagel paired with the sound of its name, the connections between the visual nodes and the auditory nodes are incremented so that, when this particular bagel recurs, it will tend to reproduce the sound of the word. But all bagels are different, and many of them get eaten, so that the exact same bagel rarely reoccurs. What happens, through learning about each bagel, is that a composite memory trace of the average bagel gets built up; the central tendency gets built up but the random variability gets cancelled out.

This property does not seem so surprising, but it coexists with another property that often does seem surprising: it turns out that multiple, unrelated associations (e.g., between the visual pattern for bagel and the sound bagel, and the visual pattern for cupcake and the sound cupcake) can coexist on the same set of connections. Experiences with different bagels and experiences with different cupcakes will all produce changes in the same set of interconnections. The resulting composite memory trace will contain the central tendencies of both associations. It is as if the traces of different exemplars of each of the different associations are averaged separately, in the sense that the mechanism allows the traces for each association to coexist. This is a very rough characterization of the state of affairs; for a fuller understanding, see McClelland and Rumelhart.[2]

Because of these properties of distributed activation models, they provide natural accounts of such things as how we learn prototypes of concepts from distorted exemplars, how we learn linguistic rules from examples that conform to them, and how semantic memories emerge from specific episodic experiences. The first point has been amply documented in the work of Anderson.[3,4] Rumelhart and I have shown how a distributed model can provide a detailed description of the acquisition of the past tense of English by young children.[5]

There are a number of other aspects of human memory data that are readily explained by distributed models. Rather than elaborate on these, however, I will consider the relations between distributed models and the brain.

RELATION TO PHYSIOLOGY

Distributed models of memory are, of course, strongly inspired by basic properties of the brain. However, it would be incorrect to imply that authors of distributed models generally mean to assume that the nodes in the models correspond exactly to real neurons, or that the connections correspond to individual physical synapses between neurons. Generally, the models should be seen as more abstract than the actual physiology, focusing instead on providing detailed accounts at the psychological level.

On the other hand, the basic tenets of distributed models seem to correspond rather closely to possible models about the physiological implementation of information processing and memory in the brain. Thus, we can hope that the basic psychological properties that emerge from distributed models will correctly characterize properties of the neural substrate of learning and memory. Further, we can hope that it will not be too difficult to specify how particular models might be mapped onto the neurophysiological substrate. This is a major advantage of distributed models over computational models based on the analogy with the traditional computer. While such models, at least in some cases, provide a detailed account for human behavior, they do not necessarily make it easy to see how the processes they specify might actually be implemented in the brain.

RELATION TO THE PHENOMENA OF AMNESIA

Assuming that we believe distributed models to provide a reasonable account of the physiological substrate of information processing and memory, the following question arises: How would we be led to think about amnesia from a distributed point of view? It is probably premature to try to give a full answer to this question, but the following three issues can be addressed. How can distributed models of memory be reconciled with basic aspects of amnesia? How can we account for the residual learning that is observed in those domains where amnesiacs show deficits? Why does this form of amnesia seem to affect some aspects of memory but not others? The following discussion considers these issues, restricting attention to those aspects of amnesia that Squire[6] has called bitemporal amnesia.

Basic Aspects of Amnesia

Bitemporal amnesia is defined by the four characteristics described below. (A fuller review is provided by Squire.[6]) (1) Insult to mesial temporal lobe structures results in correlated deficits of anterograde and retrograde amnesia. While there are some reports of dissociation of these two aspects of amnesia, it is well established in cases of amnesia due to electroconvulsive shock that anterograde and retrograde amnesia are correlated in severity; both develop gradually through repeated bouts of electroconvulsive shock.[7] (2) 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 controversially, 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.[8,9] (3) 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 years' duration based on careful experimental tests. (4) Most strikingly, memories that appear to be lost after an insult may come back. As the ability to acquire new memories recovers, 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, lasting variously from days to minutes.

At first, some of these findings seem to be difficult to reconcile with a distributed model of memory. If all memories, old and new, are stored in the same set of connections, why is it that an amnesic insult selectively disturbs the newer ones? And why is it that the older 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 stored, superimposed in the same set of connections.

However, Rumelhart and I[2] were able to provide an interpretation of these phenomena, in the context of a distributed model, based on the following assumptions. We assumed that each processing experience resulted in chemical-structural change in a large number of synapses in which many other traces had already been laid down; but that each new change underwent a gradual consolidation process, as well as a natural tendency to decay or return to the prechange state. Thus, we assumed that the changes resulting from a particular experience were widely distributed at one level of analysis, but that, at a very fine grain, within each individual synapse each change in its efficacy has a separate consolidation history. We assumed that consolidation made the residual part of the change less susceptible to decay and 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, but later decays more slowly.

So far this explanation simply takes existing consolidation accounts of the amnesic syndrome[10] 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 existing account to explain two of the important characteristics of the bitemporal amnesic syndrome. First, the hypothesis does 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 proposed that there exists a factor that we called gamma (which might be some kind of molecule) that is depleted by insult to the mesial temporal lobes. Gamma serves two functions in our model: it is necessary for consolidation (without gamma, new memory traces do not consolidate) and it is necessary for expression (without gamma, recent changes in the synapse do not alter the efficacy of the synapse, they are just ineffectual addenda, rather than effective pieces of new machinery). Implicit in these assumptions is the idea that gamma is only necessary during consolidation. Fully consolidated memories no longer need it for expression.

One helpful analogy for understanding this model is to think of a synaptic change as a new piece of functional structure, and consolidation as the glue that, when set, will hold it together and allow it to function. In this analogy, gamma is simply the clamp that holds the structure together. Without the clamp, the structure will fall apart before the glue is set; but once the setting is done, the clamp is no longer needed.

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

Space prevents a fuller discussion of this model here; suffice it to say that Rumelhart and I have completed a computer simulation of this model,[2] and we have shown how it can account for all of the aspects of bitemporal amnesia described above, simply by assuming that the amnesic insult depletes gamma, and recovery amounts to its gradual return to pre-traumatic levels. With a single parameter—gamma—that varies as a function of the amnesic insult, it accounts for phenomena ranging over a wide range of time scales; the rapid decay of information just presented to an amnesic subject over seconds and minutes, and the very extended backward reach of retrograde amnesia, over months and years.

So far I have said little about the fact that amnesiac's inability to acquire new material is not by any means completely uniform. There is now a very large literature on these spared learning effects.[6] The following summary seems to capture the basic characteristics of what is spared and what is not; a number of the papers in this volume provide reviews and viewpoints on these effects.

At one extreme, amnesiacs seem to be highly deficient in the ability to form accessible traces of particular individual episodic experiences; at another extreme, 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. In addition, they show repetition priming effects as large as those exhibited by normal subjects in a variety of perceptual and association tasks. Within the domains where learning is impaired, even the densest amnesiacs seem to learn, however gradually, from repeated experience.[11] As an everyday example if this, H.M. is clearly aware that he has a memory deficit[10] though he can recount no particular episode in which the deficit was manifest.

Residual Learning in Domains Showing Deficits

Distributed models provide a natural way of explaining why there should be residual ability to learn gradually from repeated experience even within those domains where amnesiacs 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 amnesiacs, then the general 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 repeated exemplars—automatically provide an account of the gradual accumulation of repeated information 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.

Spared Learning of Skills

However, distributed models do not account directly for the fact that certain kinds of learning are completely spared in amnesiacs, while others are not. It seems instead that the brain maintains a distinction between those synaptic connections underlying explicitly accessible episodic and semantic information on the one hand and those underlying priming phenomena and the acquisition of certain kinds of cognitive skills.[12,13] From a distributed memory point of view, we can speculate briefly on just why this might be. One conjecture is that one-trial learning of the contents of particular episodes and of specific propositions describing facts that can be explicitly accessed and considered requires very large changes in connection strengths; while the tuning of the networks that underlie cognitive skills is best carried out through very gradual modulation of connections, or at least does not require massive changes in a single trial. From this point of view, we would see the hippocampus and the consolidation processes it supports as providing a special booster to the strengths of synaptic connections in those parts of the brain over which the distributed patterns that stand for the accessible aspects of previous experience are represented.

CONCLUSION

The distributed approach to memory is exciting because it seems to do several desirable things at once. First, it provides a natural account of many aspects of information processing, learning, and memory. Second, unlike many other information-processing frameworks, it has more or less direct ties with possible neurophysiological implementations of the mechanisms it proposes. Third, it provides a relatively staightforward framework for interpreting much of what we know about certain neurophysiological deficits, such as amnesia. The approach has also been applied to account for aspects of deep dyslexia, semantic confusions of Wernicke's aphasics, and other aspects of degradation of function through brain damage.[3,14]

Work on distributed models of memory and learning is just beginning, and there is a lot more work to be done. But the models appear to provide a much more natural framework for integrating our knowledge of the neuropsychology, neurophysiology, and neurochemistry of memory with an understanding at a level closer to observable behavior.

REFERENCES

1. Hinton, G. E. & J. A. Anderson. 1981. Parallel Models of Associative Memory. Erlbaum Press. Hillsdale, NJ.
2. McClelland, J. L. & D. E. Rumelhart. 1985. Distributed memory and the representation of general and specific information. J. Exp. Psychol.: Gen. (In press.)
3. Anderson, J. A. 1983. Cognitive and psychological computation with neural models. IEEE Transactions on Systems, Man, and Cybernetics **SMC-13:** 799–815.
4. Knapp, A. & J. A. Anderson. 1985. A signal averaging model for concept formation. J. Exp. Psychol.: Human, Learning & Memory. (In press.)
5. Rumelhart, D. E. & J. L. McClelland. 1985. Acquisition of a linguistic rule by a parallel distributed processing system. *In* Parallel Distributed Processing: Explorations in the microstructure of cognition. J. L. McClelland & D. E. Rumelhart, Eds. Bradford Books. Cambridge, MA. (In press.)
6. Squire, L. R. 1982. The neuropsychology of human memory. Annu. Rev. Neurosci. **5:** 241–273.
7. Squire, L. R., P. C. Slater & P. Chace. 1975. Retrograde amnesia: temporal gradient in very long-term memory following electro-convulsive therapy. Science **187:** 77–79.
8. Huppert, F. A. & M. Piercy. 1978. Dissociation between learning and remembering in organic amnesia. Nature **275:** 317–318.
9. Squire, L. 1981. Two forms of amnesia: an analysis of forgetting. J. Neurosci. **1:** 635–640.
10. Milner, B. 1966. Amnesia following operation on the temporal lobes. *In* Amnesia. C. M. W. Whitty & O. L. Zangwill, Eds.:109–133. Butterworth & Co. London.
11. Schacter, D. 1985. Priming of old and new knowledge in amnesic patients and normal subjects. Ann. N.Y. Acad. Sci. (This volume.)
12. Cohen, N. 1985. Different memory systems underlying acquisition of procedural and declarative knowledge. Ann. N.Y. Acad. Sci. (This volume.)
13. Squire, L. R. & S. Zola-Morgan. 1985. The neuropsychology of memory: Humans and non-human primates. Ann. N.Y. Acad. Sci. (This volume.)
14. Wood, C. 1978. Variations on a theme by Lashley: Lesion experiments on the neural models of Anderson, Silverstein, Ritz and Jones. Psych. Rev. **85:** 582–591.