Memory as a Constructive Process: The Parallel Distributed Processing Approach

James L. McClelland

In Harold Pinter's play *Old Times*, a husband and wife of many years reminisce about the early days of their relationship, while awaiting a visit from the wife's best friend from that era. In these reminiscences, we learn just how differently two people can remember what were ostensibly the same events, and ostensibly the same people – most notably, themselves and the wife's best friend. For each, these reminiscences have become embedded in a complex, not fully consistent, and self-serving personal history that does not survive well when juxtaposed against the reminiscences of the other.

In an interview at the time of the opening of *Old Times*, Pinter was asked to comment on his thoughts while writing his play. He said "what fascinates me is the mistiness of memory"(1). Cloud-like, forever changing, memories are clearly not like frozen snapshots taken on a day long ago, pulled out from the back of a drawer for re-inspection. Indeed, since the work of Frederick Bartlett in 1932, memory researchers have been keenly aware of the constructive nature of memory (2). Bartlett asked educated people at Cambridge to read, and then later to recall, a story from a native North American culture. Though written in English, the story had an unfamiliar structure and content. The recollections of the participants retained elements from the original story but many details were omitted or transformed in ways that seemed to Bartlett to fit better with the cultural context of the individuals who were recalling them. Repeated attempts at recall by the same individual resulted in gradual fixing of the elements, but into a story sometimes quite different from the original. Such findings led Bartlett and others to view recollection as a process not unlike the activities of an archaeologist faced with the task of reconstructing an ancient dinosaur from a collection of bones found near each other. The ultimate product contains some of the fragments of one, but possibly not only one, dinosaur, and many parts are filled in based on the archaeologists' knowledge of other similar dinosaurs. Close resemblance to any real creature that once lived is far from guaranteed.

The idea of memory as a constructive process provides a bridge between the worlds of art and science, since the idea clearly has its protagonists in both spheres. This chapter offers a scientific theory of the nature of human memory that fits very naturally with this constructive perspective. The theory, which we call the complementary learning systems theory, developed in three stages. The groundwork for this theory was laid during the development of a broad framework for understanding human cognitive processes called the Parallel Distributed Processing framework, a project I participated in with David Rumelhart and others in 1986. (3). The subsequent development of the theory itself occurred in the early 1990's and was presented in a 1995 paper by Bruce McNaughton, Randall O'Reilly and myself (4). At that time, the focus was on one of the two complementary learning systems, a fast learning system in the medial temporal lobes of the brain. Subsequent work with Tim Rogers, presented in our 2004 book, *Semantic Cognition*, focused on the other, gradual developmental learning system located in other regions of the neocortex (5). It should be noted that the theory is not universally accepted and still has many gaps; furthermore, some important recent discoveries have not been fully integrated into it. This chapter introduces the groundwork for the theory, distinguishes it from other researchers' approaches to the neuroscience of memory, lays out the theory itself, and considers recent developments. The final section opens questions for the theory and gives some final thoughts on its relevance to the arts and humanities.

Neurons and Synapses: The Physical Substrate for Representation and Memory

The complementary learning systems theory is grounded in a way of thinking about representation and memory in the brain that arose in the 1980s when during the development of the PDP framework. The starting places are the crucial physiological building blocks of the neocortex of the brain: neurons and synapses. The human brain contains nearly 100 billion neurons, and each neuron has from 1,000 to 100,000 synapses: points of contact with other neurons. Figure 1, a famous drawing by the 19th century neuroanatomist Ramón y Cajal, evokes a sense of the actual physiology (6). This is a drawing Ramón y Cajal made from what he could see through a microscope in a very thin slice of brain tissue that he had treated. The treatment he used caused one out of every 100 neurons in the slice to turn black, allowing the viewer to visualize the structure of individual neurons. Each neuron is a cell, with its own cell body – the pyramid-like blobs in the figure. Coming out of each cell body there are several branching structures – the dendrites, the heavier branches with tiny bumps along them that reach up and branch to the sides of the cell body – and the cell's narrow axon, which arises from the bottom of the little pyramid-shaped body and projects downward, with branches that turn back up into the tissue. What the figure does not show is the dense branching of these axons into tiny filaments and the terminals of these axons on the dendrites. However, the little bumps along the dendrites, called synaptic boutons, are the main locations where these connections are made. Envision, if you can, this structure in its full three-dimensional splendor, with dendrites and axons branching out to the front and to the back as well as to the sides, and with 100 times as many neurons packed into the same space. This is just one square millimeter of the human neocortex, only about 1 millionth of the entire cortical volume.

Neurons and synapses constitute the physical substrate for our active mental states and our memories. An active mental state can arise from perception – say of the sound of a person's name, or the sight of a person's face – or from thinking – as when, for example, we see a cat creeping up on a bird and have the thought that the bird may fly away. In our theory, these active mental states are patterns of activation over populations of neurons across many regions of the neocortex of our brain. This raises two further points: the localization of different aspects of mental content in different brain regions and within brain regions, and the question of whether content is localized in individual neurons.

With the aid of Figure 2, we can address the first point. This figure shows two views of the left hemisphere of a typical brain. One viewed from the side (as if one is looking through the skull of a person one is looking at, when that person is looking to the left), and one viewed from below (as though the hemisphere were tilted away from the viewer and laid on its flat, inner side). Most of the colored regions illustrate areas that become active when a person perceives, or is asked to bring to mind, a particular kind of information about an object. In one valuable experiment, participants were shown words like 'pencil' and were asked to think of the color of the object denoted, or of the action one performs on the object. In the first case, activation was found in the region labeled *color* and in the second, in the region labeled *action*. Other studies have contributed evidence relevant to the other regions illustrated in the figure. There are also specialized areas for information about faces, and other specialized areas for words, capturing their sound, their articulation, and their spelling. Thus, it seems fairly well established that different kinds of information are represented in different parts of the brain.

Localist vs. Distributed Representation. The role of individual neurons within each brain region is less settled. One view is that individual neurons stand for entities, or properties of entities we recognize intuitively and can easily label or describe. Representations of this type are often called localist representations, because you can locate within the representation the neurons that capture particular aspects of the information, and they are sometimes also called 'grandmother cell' theories, because, among other things, they propose that individual neurons represent specific familiar objects such as one's grandmother. In this view, upon seeing grandmother, some active neurons have the specialized role of representing the crinkly lines around her eyes, others the grayish tint of her hair, and still other neurons have the specialized role of representing grandmother. This is an easy theory both for the scientist and the layman to grasp, and is still widely discussed (7).

The alternative to the grandmother neuron concept is the concept of distributed representation. The idea has a long history, and was set forth within the PDP framework by Geoff Hinton, David Rumelhart, and myself. (8). In distributed representations, the focus is on the pattern of activation as a whole, and not on the individual neuron. In this view, each individual neuron participates in the representation of many things different things, and no neurons are dedicated to individual items. Whether these things that activate an individual neuron all share something that can be labeled or described is not of the essence here; what is essential is that the representations of things that are similar (in terms of the kind of information represented, e.g., the action one takes on the object for example) involve highly overlapping populations of neurons. Thus, in terms of neurons representing the visual appearance of a cheetah, a leopard, and a flamingo, the pattern for the visual appearance of the cheetah and the leopard will have far more units in common than either has with the pattern for the flamingo. The similarities in question can be of many different types, including similarities in such abstract domains as professions. Therefore, in the appropriate brain region, comedians will be represented by patterns that overlap more with each other than they do with the patterns associated with politicians.

An Integrative Representation Independent of Any Specific Kind of Content?   A further feature of our theory is that there should be, somewhere in the neocortex, an area where there is an integrative representation of all sorts of things, encompassing all aspects of their content (9). A number of other theorists have proposed related ideas (10). There are still alternative perspectives, however (11). Research is ongoing on this issue, but some evidence points to the possibility that such a representation may exist in the anterior temporal cortex, sometimes called the “temporal pole”, and labeled as such in the figure. According to our theory, whether one hears the word “dog,” or hears a dog barking, or sees a dog, or thinks about how a dog responds when greeting one’s houseguests, a representation becomes active in the anterior temporal cortex. Each particular thought of a particular dog will evoke a slightly different representation, but the representation of any dog will generally have much more overlap with the representation of a goat, say, than it will with the representation of say, a maple tree.

The Knowledge is in the Connections:
Acquisition of Semantic Memory Through Connection Adjustment

When a speaker produces the word *dog* vibrations reach the ear, giving rise to the firing of neurons in the auditory pathway. How can this give rise to a pattern of activation corresponding to the typical color of a dog or of the sound of the dog barking, or of the way the dog wags its tail? According to our theory, this depends on the pattern of interconnection among neurons. Connections carry signals from neurons in the ear to neurons in primary auditory cortex, and from there to neurons in higher auditory cortex. Perhaps through further intermediaries, connections then carry signals to the neurons that participate in the integrative representation, and still other connections carry signals from these to neurons representing each of the different kinds of content. The connections in the auditory pathway itself are initialized early in development before the eyes open, and for our present purposes, we can treat these connections as if they were fixed. But how do we activate the neural pattern for the color, shape, sounds, and movements of a dog from a higher-level visual representation of neural representation of the spoken word "dog"? Clearly, knowledge acquired from experience is necessary, since the relationship between the word and the objects that it stands for are idiosyncratic and language specific. For this reason, many researchers treat this knowledge as a form of memory – often called "semantic memory." In our theory, memory of this kind and memories of other kinds are stored in connections. The idea that knowledge and memory is stored in connections is sometimes called *connectionism,* and theories based on this idea are often called *connectionist theories*.

A schematic illustration of this concept is shown in Figure 3. Let's imagine, for concreteness, that we are considering connections that allow a pattern corresponding to the sound of the name of an object to produce a pattern corresponding to one other kind of information -- say, a pattern representing what the object looks like. This is a simplification of the theory introduced above, because there we noted that in the theory the integrative representation actually mediates between the name and the other kinds of information. At the outset of learning we imagine that we have very weak, non-specific synaptic connections between the neurons in the auditory representation and those in the visual representation. In this situation, the activation of the pattern for the sound of the word, followed closely by the activation of the pattern for the appearance of the object, creates the conditions needed for strengthening the connections from the neurons active in the sound representation to the neurons active in the auditory representation.

The idea that pre-, then post-synaptic activation will lead to the strengthening of connections among neurons is a variant of the famous proposal of Donald Hebb in 1949 (12), and is widely discussed in other chapters of this book. Hebb established a starting point for a large number of experimental investigations and computational models of the underlying physical process that provides the substrate for learning and memory. It should be noted that the exact formulation of the details of the 'synaptic modification rule' is a subject of considerable ongoing investigation both in computational as well as experimental investigations.

It seems fair to say that we still don’t fully understand exactly how the brain achieves its remarkable success in making connection adjustments that successfully form the substrate of learning and memory.   We do assume that the brain can do so even when intermediate or “hidden” neurons, like those in our integrative layer, are involved.   Such a network is illustrated schematically in Figure 4. Inputs of different kinds specify the patterns of activation representing different kinds of information about an item such as a dog, but do not specify what pattern should be used for the item’s representation on the integrative layer. Using a sophisticated connection adjustment rule (13), it is possible for repeated experience with many different things to produce cumulative adjustments to the connections. The presentation of any unique aspect of one of these known things (the aroma, or the prick, or a rose; the bark of a dog or the spoken word *dog*) will give rise to activation of an item specific pattern on the hidden layer and of the appropriate item specific patterns across all of the visible layers (14).

What is the Memory Trace of an Experience? What is most important in the present context is not the exact nature of the connection adjustment rule that is used, but the more fundamental fact that the memory trace left behind by a specific experience is a pattern of connection adjustments.   This theory is very different from the standard notion that the memory trace of an experience is viewed as a record of the experience itself, somewhat like a memorandum (lacking details perhaps) that can be filed away in a drawer for subsequent retrieval. One key difference is that, in our theory, the memory traces of different experiences are not kept separate. In the case of a network that learns about words and the objects they describe, each objects name will be a pattern of activation that overlaps with the patterns for the names of other objects, and each objects visual representation will be a pattern that overlaps with the patterns corresponding to the visual representations of other objects.   Repeated experiences in which one hears the word “dog” and sees a dog will gradually lead to the buildup of strong connections allowing the word to activate the visual pattern. Repeated experiences with other similar objects or objects with similar names will, also, affect the same connections.

According to our theory, then, there is no possibility of retrieving a specific memory. A memory does not exist in its own separate storage location – its residue in the brain is distributed over many synaptic connections, whose values have also been shaped by many other experiences.   Thus, for example, remembering what a dog looks like upon hearing the word “dog” is always a constructive process, one that involves the participation of influences arising from many experiences overlapping with each other in various ways and in various degrees.   In a nutshell: remembering in a system that uses connection adjustment between neurons participating in distributed representation is intrinsically a constructive process. Let us now consider some important facts that have played a key role in shaping the development of this complementary learning systems theory.

Key Discoveries about the Brain Basis of Memory

Our understanding of the brain basis of memory took a quantum leap forward after the surgeon William Scoville removed both the left and right medial temporal lobe from the patient HM to treat his intractable epilepsy (15). Upon waking up after the surgery HM recognized family members and could converse apparently normally. In formal testing, his IQ was in the normal range, and he performed at least as well as he had previously, and within the normal range, on tests of general knowledge, vocabulary, and on memory for the early periods of his life. He could carry out attention demanding tasks at normal levels. As is well known, however, HM exhibited profound and very striking deficits. Most obvious was the fact that he could not form new memories either of people or events. A person not previously known to HM could come into his hospital room and carry on a conversation with him for any length of time; if the person then left the room even for a minute or two and came back in again, HM would not recognize him or remember the conversation they had been engaged in only moments before. Thus, HM had a profound deficit in the ability to form new memories, although he retained a great deal of knowledge he had acquired before the surgery.

The profound loss of the ability to form new memories seen in HM has now been documented in many other patients with similar patterns of brain damage, and there are some other cases where the loss is even more profound. Further studies of HM and many other patients also underscore three additional important points (16).

Normal Acquisition of Skills. While some forms of memory were profoundly impaired, other forms were not. In one study with HM, he was repeated asked to trace figures while viewing the figure and his hand in a mirror. Like others, he was initially very bad at this, often making movements in the wrong direction. With practice he improved, however, and he appeared to improve at about the same rate as healthy normal individuals. This was so, even though he never had a recollection of having performed the task before. Similar findings have been reported in a large number of other studies.

Normal Levels of Item-specific Priming. Another form of memory that appears to be spared in patients with Medial Temporal Lobe lesions is revealed when the patient is tested for subtle after-effects of previous experiences with individual items, such as words or pictures. An example from the studies with HM involved a standardized picture fragments task. In this task, the patient is shown several series of cards. The cards in each series contain an increasing number of more fragments or line segments from a drawing of a familiar object, such as an airplane. Thus, if HM needed 70% of the fragments to recognize the airplane the first time through the airplane series, he might only need 50% of the fragments the second time. Importantly, the size of this improvement is about the same as that seen in normal subjects, even though the patient will not recall having seen a series of cards with fragmentary drawings of an airplane before.

Graded Loss of Memories from Experiences before Surgery. As time goes on after an experience, the memory for it becomes less susceptible to the effect of hippocampal removal HM and other patients appear to have no memory for events occurring within a period of months, or even years, before the surgery. Memory for events from early life may appear to be intact. For example, HM had no recollection after the surgery of ever meeting Dr. Scoville prior to his surgery, nor did he recall that he had consented to the operation, although he could recall many events from earlier periods of his life. There is disagreement about how far back the retrograde amnesia can extend. The phenomenon is difficult to study in humans because it is difficult to document each individual’s prior experiences clearly enough to assess how well they are remembered. There is evidence that, in some patients the deficit may extend over several decades (17).

The Complementary Learning Systems Theory

Several different explanatory frameworks can help make sense of these findings. Some of these frameworks rely on the idea that there are many separate memory systems in the brain. One notable version of this view by Larry Squire in 1992 is that these memory systems are divided into two types – declarative and non-declarative (16). Declarative memories are memories we attest to, while non-declarative memories are memories that somehow affect our behavior, even when we may be unaware of the experience in which these memories were formed.   Squire summarizes the facts by proposing that the MTL region is involved in the formation of new explicit memories and the recall of recent declarative memories, but it is not involved in the formation of or recall of non-declarative memories. Squire, following Brenda Millner (18) and others, also suggested that some unspecified process often labeled *consolidation* occurs after memories are first acquired, such that, over time, they become independent of the medial temporal lobe memory system.

The theory that my colleagues and I developed attempts to go beyond this level of description (4). Drawing ideas from an earlier theory of David Marr (19), it captures in more detail the mechanisms involved in the formation and retrieval of all kinds of memories, and it provides one way of understanding why it makes sense to have more than one learning system involved in memory formation.

According to our theory, when someone processes an item – perhaps a fragmentary picture of an airplane – patterns of activation arise in early stages of the visual processing stream, leading up to the visual association areas when the shape or form of the pictured object is represented. If enough fragments are presented, the visual pattern will be close enough to that of previously seen airplanes to give rise to patterns of activation in areas representing many of the different types of information about an airplane, including the pattern corresponding to the spoken word “airplane.” Small adjustments to the strengths of the participating connections will then occur. The consequence of these small adjustments is to slightly facilitate the subsequent processing of the same item, and to make it possible for the system to enter the distributed state of having recognized the airplane with slightly fewer fragments than before. In this way, our model addresses the subtle after-effects of specific episodes of processing.

An essential element of our complementary systems theory is that the connection adjustments are very small and have only subtle after effects. Such adjustments may cumulate over repeated presentations of the same input and corresponding output (for example, the reflected letter R in reversed presentations of many different words containing this letter), so that gradually the ability to read the reversed letter R correctly will build up. But, according to the theory, these connection adjustments are not large enough to allow an arbitrary new association – say between a person’s name and his face – to be formed after one or even a few exposures.

In order to allow the rapid formation of such new associations, we proposed that the medial temporal lobes provide a learning system that complements the gradual learning system in the neocortex. This idea is sketched in Figure 4. Here we see the MTL region sending and receiving connections to and from all of the important representational areas in the neocortex. When a person sees someone and hear his name, patterns of activation arise in the relevant cortical areas. Patterns arise in other areas as well, corresponding to the location in which the person and the name is encountered, the emotional state associated with the encounter, the sound of the person's voice, etc. These in turn are propagated into the medial temporal area via the axons of neurons arising within each of the involved areas. These inputs then set up a pattern in several regions of the hippocampus, deep inside the medial temporal area, corresponding to the entire experience of seeing the person and hearing his name. Large connection adjustments then occur among the neurons participating in the hippocampal representation, which have the effect of binding together the elements of the hippocampal representation of the encounter so that, if a part of the input is presented again at a later time – for example, one sees the person's face a week later -- the pattern will tend to be reconstructed . Return connections from the hippocampus to the contributing neocortical areas then allow the corresponding cortical pattern to be reconstructed. This pattern, in our theory, corresponds to the experienced memory for the previous event.

It should be clear how our theory can explain why a patient like HM would fail to form new arbitrary associations between a new person's name and face and would also fail to associate these things with the episode in which the individual was encountered. In HM's case, almost the entire medial temporal area, including the hippocampus, was removed on both sides of his brain. Thus, the neural substrate for forming such new traces would largely be absent from the system. Interestingly, the theory explains why HM would fail to remember a person or episode encountered shortly before his surgery. The physical substrate of the memory for the episode – the brain areas containing the neurons and connections in which the memory trace of the experience was stored – would no longer be available to play a role in the construction of the distributed pattern over the neocortical brain areas that would correspond to the (re)constructed memory.

But why, within this theory, should a memory gradually become less dependent on the presence of the medial temporal area with the passage of time? While the memory trace resides in the hippocampus, events occur that can trigger reinstatement. These events might correspond to waking experiences which include encountering a same person again, causing a previous experience to come back to mind. They may correspond to conscious and deliberate recollection of the previous experience; or they may correspond to spontaneous reactivations of hippocampal patterns from memories during sleep or even during waking moments. Accordingly each such reactivation would provide the slow-learning neocortical system with another chance to learn. Therefore, gradually, over many such repetitions, connections within the neocortex would be strong enough to allow the person's face to give rise to the pattern corresponding to his name.

In short, the complementary learning systems proposed in our theory – one in the neocortex, which relies on small connection adjustments, and one in the medial temporal lobe, which relies on very large connection adjustments – work together, over time, to provide an overall result that allows declarative memories to become consolidated, gradually losing their dependence on the medial temporal lobes.

Cooperation of Complementary Learning Systems in Memory for Meaningful Materials. According to our complementary learning systems theory the neocortex and the medial temporal lobes generally work together in remembering. Associations that already have some pre-existing strength (e.g. between the word 'dog' and the word 'bone') will benefit from the synergistic contributions of both systems, while those that are completely arbitrary (e.g. city – ostrich) will depend almost entirely on the fast-learning medial temporal lobe system. The stronger the pre-existing strength of the connections in the cortex, the less important the medial temporal lobe contribution will be. Other studies have shown that pre-existing associations can also contribute to false recollection, so that if ‘dog’ and ‘bone’ occur in different sentences one has heard, the words and other words they occurred with in the same sentence may sometimes be recalled together (20). Here the cortex and MTL are working at cross-purposes, creating a false memory.

Recapitulation: Cooperation of Complementary Learning Systems in the Repeated Recall of an Initially Unfamiliar Story. We can now go back to considering something as complex as the successive recollection of the story Frederick Bartlett presented to the subjects in his experiment. According to our theory, these subjects initially made connection weight adjustments within their medial temporal lobes at the time of their initial reading of the story. Processes operating at the time of reading the story in the first place may well have distorted their understanding as they took the story in, and although we have not discussed this extensively above, such things are very much expected to happen within the overall Parallel Distributed Processing framework (21). The patterns of activation arising in the neocortex in the course of the reading of the story would then give rise to patterns in the hippocampus, and fast changes in the strengths of connections among the participating neurons would in turn provide the initial memory trace that contributes to the participant’s effort to reconstruct the story at a later time. This process, however, would also be affected by prior associations of recollected elements of the story with other things known to the participant, giving rise to the opportunity for selective memory for elements of the story that make sense and distortion in the recollection in the direction of further sense-making. This act of reconstruction would lead to further connections within the medial temporal lobes, as well as to small adjustments to relevant connections within the neocortex. If the process of recollection were repeated often, we would expect a gradual strengthening and increase in the consistency of recall over time, along with a gradual reduction in dependence on the medial temporal cortex.

Questions for the Complementary Learning Systems Theory

We have described evidence that there are two complementary learning systems, but we have not yet explained why this would be desirable or necessary. Specifically, we specified that the neocortex makes only small connection adjustments, insufficient to store new arbitrary associations rapidly. But why should this be so? Why shouldn't the size of connection adjustments in the neocortical system be increased so as to allow new information to be stored rapidly? This question lies at the heart of the 1995 article in which we presented the theory (4). There we showed two important things:

First, the ability to generalize what we learn about one thing to other things and to find the statistical regularities underlying a range of related experiences depends on slow learning. If one makes very large connection adjustments, the idiosyncratic aspects of particular experiences dominate learning too much, and the connection weights fail to capture the common structure underlying a set of related events and experiences.

Secondly, if one has gradually built up a body of knowledge in a neocortex-like slow-learning system, any attempt to add arbitrary new information into the connection weights at all once will create a phenomenon known as ”catastrophic interference” (22). This means that although the new learning may be possible, it will drastically interfere with what has previously been stored in the system. For example, if one forces a network to learn about a penguin – a bird that can swim but cannot fly – this information can be learned, but it interferes with pre-existing knowledge about what other birds can do. Crucially, though, one can overcome catastrophic interference, if one engineers the situation correctly. If one interleaves presentation of the new information about the penguin with ongoing exposure to information about other birds, knowledge that the penguin is a bird that can swim but cannot fly is gradually learned, and about what other birds can do is maintained.

This example points out why we have complementary learning systems. The fast learning system provides a way to store and remember new arbitrary information quickly but in a separate system from the one containing our pre-existing knowledge about other things. Once stored in the fast learning system, this information can be replayed occasionally, interspersed with replay or ongoing experience with information about other things. Gradual learning of the new information then occurs in the cortex, without catastrophic interference. The fast learning MTL system, working together with the neocortical system, thus provides a way to eventually knit the newly formed memory into the fabric of what is already known to the slow-learning neocortical system.

Does Neuroscientific Evidence Support the Basic Tenets of the Complementary Systems Learning Theory? When we initially developed the theory, we sought evidence that would support or refute it in a number of different places. First, we asked whether the necessary long-distance pathways exist in the brain, to carry out the necessary long-range interactions with the neocortex. Indeed, evidence collected by the neuroanatomist David Amaral and his collaborators was completely consistent with the theories' requirements (23). In brief, there are two-way fiber bundles connecting the hippocampus with all relevant areas of the neocortex and with other relevant non-cortical brain areas. Second, we found supporting evidence for the differences in synaptic plasticity between hippocampus and neocortex. Indeed, the phenomenon of *long-term potentiation* (LTP) – in which simultaneous pre- and post-synaptic activation gave rise to long-lasting changes in strengths of synaptic connections – was first described in slices taken from the hippocampus (24). While LTP can be produced in the neocortex, a study by Ronald Racine and his collaborators showed, as the theory predicted, that hippocampal LTP reaches maximum levels quickly, while in neocortical synapses changes are very small each time the stimulation is applied and build up gradually with repeated exposures (25). Finally, we asked whether in fact there was evidence of reactivation of patterns of activation established during a learning event at later times, in particular while an animal was sleeping. In 1992, when we were first developing the theory, there was little known about the matter, but a study during that time by Matthew Wilson & Bruce McNaughton (26) provided the first clear support for the sleep-reactivation idea, and it is now a well-studied phenomenon (27).

Moreover, there is a vast body of research both on the role of the medial temporal lobes in learning and memory and on the biological processes underlying learning and memory. One line of evidence that has played an important role is helping to explain the exact nature of the role of the hippocampus in memory formation relates to the following question: If connection adjustment is involved in learning in the hippocampus, as well as in the neocortex, why is interference only a problem in the neocortex? Specifically, why are we able to learn new arbitrary things in the hippocampus without this new learning interfering catastrophically with other information already stored in connections between hippocampal neurons? Since connection adjustment is assumed to be involved in both the hippocampus and the neocortex, why doesn't learning something new rapidly also produce catastrophic interference in the hippocampus? An answer to this question is suggested by observing differences between the activation of neurons in both the hippocampus and neocortex in response to the same experience. Experiments from the McNaughton lab have shown that a far smaller fraction of the neurons are active at any one time in the hippocampus than in many regions of the neocortex (28). This sparser pattern of activation tends to reduce the extent to which memories for the different things are stored in the same connections, drastically reducing the amount of interference in memory. A detailed theory of how sparse representations minimize interference has been developed with contributions from a number of memory researchers, (29).

It should be noted that there are complementary benefits to the use of overlapping representations: What is learned or remembered about one thing does transfer to others. To the extent that overlap captures important elements of similarity that support such generalizations, a high degree of overlap of representations can be a very good thing. There remains debate about the degree of similarity-based overlap in neocortical representations, but the overlap is clearly greater in many areas of the cortex than in the hippocampus.

Can the Theory Address Recent Discoveries about the Neural Basis of Memory?

The complementary learning systems theory is fifteen years old and there have certainly been many new developments in the memory research. Some of these developments provide striking confirmation of details of the mechanisms we and others have proposed, such as a mechanism within the hippocampus for assigning separate, non-overlapping patterns of activation to experiences of very similar (30).

One intriguing development that was not anticipated in our theory is the phenomenon of *reconsolidation* (31). This is the finding that memories thought to have already been consolidated can sometimes be put back into a fragile state, if brought back to mind by a cue or reminder of the remembered experience. This intriguing notion suggests that memories can sometimes be erased, or possibly edited, with new information replacing information previously consolidated due to a single reminding / revision episode. If such a process could occur for all information stored in the neocortical learning system, it would pose a severe challenge to our theory. However, the generality of the phenomenon remains unclear, and there have been several failures to reproduce the effect. Also, it appears that it may be best to think of reconsolidation occurring, if it occurs at all, when memories are still relatively new and unconsolidated; those that have undergone repeated reinstatement appear to be more robust, perhaps more completely consolidated, so that returning them to a more labile state is no longer possible (32). This pattern is more compatible with the complementary learning systems theory.

It should be emphasized, however, that within the complementary learning systems theory, the overt recollection of an event can certainly be affected by the content of subsequent experience. This can occur because the adjustments previously made to the strengths of connections can have new adjustments overlaid on top of them, pushing the connections in different directions and affecting the representation of an earlier event that is reconstructed at the time of remembering.

In conclusion, the two main points of this chapter are these. First, that memory is a constructive process; and second, that the biological substrate of memory is the pattern of adjustments an experience produces to connections among neurons. The chapter has discussed how these ideas are compatible, and has shown how apparently different forms of knowledge or memory may depend in different ways of two complementary learning systems. Both rely on connections to store memory and knowledge; but their different characteristics allow them to perform different roles in our human ability to remember, and, as the playwright Pinter reminded us, to misremember.

Notes

[1] Mel Gussow, *Conversations with Pinter* (London: Nick Hearn Books, 1994), 16.

[2] Frederick C. Bartlett, *Remembering: A Study in Experimental and Social Psychology* (London: Cambridge UP, 1932).

[3] David E. Rumelhart, James L. McClelland, and the PDP research group, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume I & Volume II* (Cambridge, MA: MIT Press, 1986).

[4] James L. McClelland, Bruce L. McNaughton, and Randall C. O'Reilly, "Why There Are Complementary Learning Systems in the Hippocampus and Neocortex: Insights from the Successes and Failures of Connectionist Models of Learning and Memory," *Psychological Review,* 102 (1995): 419-457.

[5] Timothy T. Rogers and J. L. McClelland, *Semantic Cognition: A Parallel Distributed Processing Approach* (Cambridge, MA: MIT Press, 2004).

[6] Santiago Ramón y Cajal, *Comparative Study of the Sensory Areas of the Human Cortex* (Worcester, MA: Clark University, 1899), 325.

[7] Jeffrey S. Bowers, "On the Biological Plausibility of Grandmother Cells: Implications for Neural Network Theories in Psychology and Neuroscience," *Psychological Review*, 116 (2009): 220–251. Stephen Waydo et al., "Sparse Representation in the Human Medial Temporal Lobe," *Journal of Neuroscience,* 26 (2006): 10232-10234. David C. Plaut and J. L. McClelland, "Locating Object Knowledge in the Brain: A Critique of

Bowers' (2009) Attempt to Revive the Grandmother Cell Hypothesis,” *Psychological Review*: (in press).

8. Geoffrey E. Hinton, J. L. McClelland, and David E. Rumelhart, “Distributed Representations,” in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, 1: 77-109. J. L. McClelland and D. E. Rumelhart, “Distributed Memory and the Representation of General and Specific Information,” *Journal of Experimental Psychology: General*, 114 (1985): 159-188.

9. J. L. McClelland and Timothy T. Rogers, “The Parallel Distributed Processing Approach to Semantic Cognition,” *Nature Reviews Neuroscience,* 4 (2003): 310-322.

10. Antonio R. Damasio, “The Brain Binds Entities and Events by Multiregional Activation from Convergence Zones,” *Neural Computation*, 1 (1989): 123-132.

11. Alex Martin and Linda L. Chao, “Semantic Memory in the Brain: Structure and Processes,” *Current Opinion in Neurobiology*, 11 (2001): 194-201.

12. Donald O. Hebb, *The Organization of Behavior* (New York: Wiley, 1949).

13. D. E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams, “Learning Representations by Backpropagating Errors,” *Nature*, 323 (1986): 533-536**.**

14. Timothy T. Rogers et al., “The Structure and Deterioration of Semantic Memory: A Neuropsychological and computational investigation,” *Psychological Review*, 111 (2004): 205-235.

15. William B. Scoville and Brenda Milner, “Loss of Recent Memory after Bilateral Hippocampal Lesions,” *Journal of Neurology, Neurosurgery, and Psychiatry*, 20 (1957): 11-21.

[16] Larry R. Squire, “Memory and the Hippocampus: A Synthesis from Findings with Rats, Monkeys, and Humans,” *Psychological Review*, 99 (1992): 195-231.

[17] Dean F. McKinnon and Larry R Squire, “Autobiographical memory and amnesia”. *Psychobiology,* 17 (1989): 247-256.

[18] Brenda Milner, “Amnesia Following Operation on the Temporal Lobe,” in *Amnesia*, ed. Charles. W. M. Whitty and Oliver. L. Zangwill (London: Butterworth, 1966), 109-133.

[19] David Marr, “Simple memory: A theory for archicortex”. *Philosophical Transactions of the Royal Society of London. Series B*, 262 (1971): 2381.

[20] Cynthia Henderson and James L. McClelland, “Semantic Interference During Episodic Recall,” in *Proceedings of the Thirty-First Annual Conference of the Cognitive Science Society.*,, ed. Niels Taatgen and Hedderik van Rijn (Cognitive Science Society, http://cognitivesciencesociety.org/, 2009), 3203.

[21] David E. Rumelhart et al., “Parallel Distributed Processing Models of Schemata and Sequential Thought Processes,” in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume II,* ed. J. L. McClelland, D. E. Rumelhart, and the PDP research group (Cambridge, MA: MIT Press, 1986), ch. 14.

[22] Michael McCloskey and Neal J. Cohen, “Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem,” in *The Psychology of Learning and Motivation,* ed. Gordon H. Bower (New York: Academic Press, 1989), 24, 109-165.

[23] Larry R. Squire, Arthur P. Shimamura, and David G. Amaral, “Memory and the Hippocampus,” in *Neural Models of Plasticity: Experimental and Theoretical*

Approaches, ed. John H. Byrne and William 0. Berry (New York: Academic Press, 1989), 208-239.

24. Tim V.P. Bliss, and Terje Lomo, "Long-lasting Potentiation of Synaptic Transmission in the Dentate Area of the Anaesthetized Rabbit Following Stimulation of the Perforant Path," *Journal of Physiology (London),* 232 (1973): 331-356.

25. Ronald J. Racine et al., "Post-activation Potentiation in the Neocortex. IV. Multiple Sessions Required for Induction of Long-term Potentiation in the Chronic Preparation," *Brain Research,* 702 (1995): 87-93.

26. Matthew A. Wilson and Bruce L. McNaughton, "Reactivation of Hippocampal Ensemble Memories during Sleep," *Science,* 265 (1994): 676-679.

27. See Robert Stickgold, "Memory in Sleep and Dreams: The Construction of Meaning," ch. 4 of the current volume.

28. Carol A. Barnes et al., "Comparison of Spatial and Temporal Characteristics of Neuronal Activity in Sequential Stages of Hippocampal Processing," *Progress in Brain Research,* 83 (1990): 287-300.

29. David Marr, "A Theory of Cerebellar Cortex," *Journal of Physiology (London),* 202 (1969): 437-470. Bruce L. McNaughton and Richard G. M. Morris, "Hippocampal Synaptic Enhancement and Information Storage within a Distributed Memory System," *Trends in Neurosciences,* 10 (1987): 408-415. Randall C. O'Reilly and James L. McClelland, "Hippocampal Conjunctive Encoding, Storage, and Recall: Avoiding a Tradeoff," *Hippocampus,* 4 (1994): 661-682.

30. Jill K. Leutgeb et al., "Pattern Separation in the Dentate Gyrus and CA3 of the Hippocampus," *Science,* 315 (2007): 961 – 966. Thomas J. McHugh et al., "Dentate

Gyrus NMDA Receptors Mediate Rapid Pattern Separation in the Hippocampal Network," *Science*, 317 (2007): 94 – 99.

31. Karim Nader, Glenn E. Schafe, and Joseph E. Le Doux, "Fear Memories Require Protein Synthesis in the Amygdala for Reconsolidation after Retrieval," *Nature*, 406 (2000): 722–726.

32. Yadin Dudai, "Reconsolidation: the advantage of being refocused". *Current Opinion in Neurobiology*,16, (2006): 174-178.

Recommended Reading List

McClelland, James. L., Bruce L. McNaughton, and Randall C. O'Reilly. “Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory.” *Psychological Review,* (1995) 102: 419-457.

Milner, Brenda. “Amnesia following operation on the temporal lobe.” In *Amnesia,* edited by C. W. M. Whitty and Oliver L. Zangwill, 109-133. London: Butterworth, 1966.

Rumelhart, David E., James L. McClelland, and the PDP research group. *Parallel distributed processing: Explorations in the microstructure of cognition. Volume I & Volume II.* Cambridge, MA: MIT Press, 1986.

Rogers, Timothy. T. and James L. McClelland. *Semantic Cognition: A Parallel Distributed Processing Approach.* Cambridge, MA: MIT Press, 2004.

Squire, Larry R. “Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans,” *Psychological Review,* 99 (1992): 195-231.

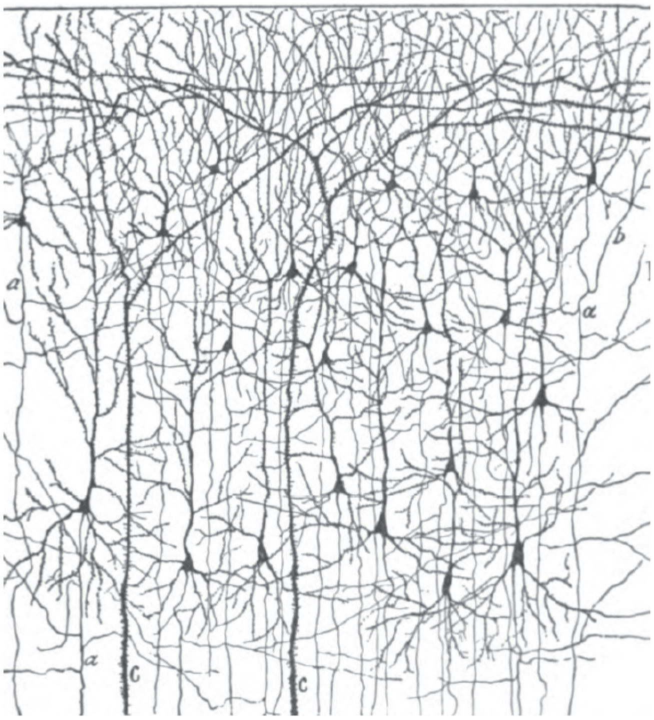

Figure 1.

The microanatomy of the cerebral cortex, as drawn by the Spanish neuroanatomist Santiago Ramón y Cajal. From Santiago Ramón y Cajal, *Comparative Study of the Sensory Areas of the Human Cortex*(1899), 325. Permission Pending.

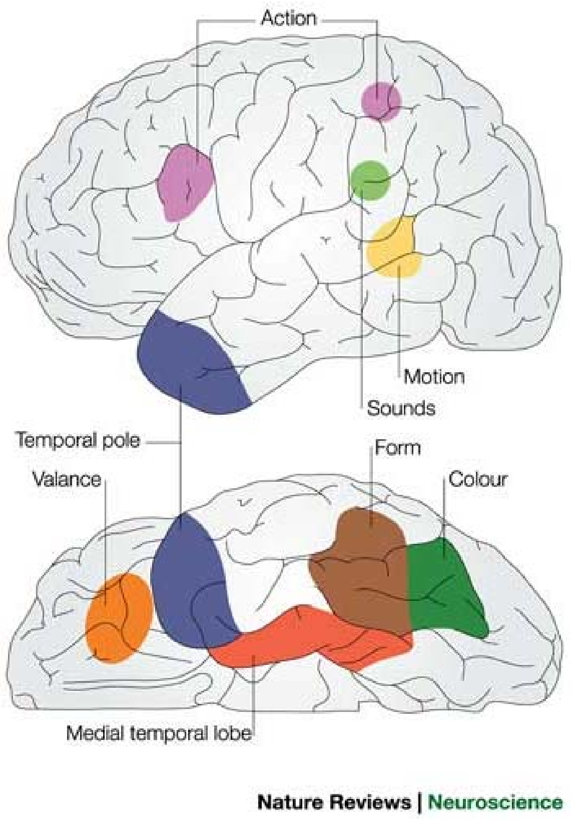

Figure 2

.

Lateral view (top) and view from below (bottom) of the left hemisphere of the human brain, showing areas representing that become active when a person processes different kinds of information about an object. From J. L. McClelland & T.T. Rogers," The Parallel Distributed Processing Approach to Semantic Cognition" (2003),315. Permission Pending.

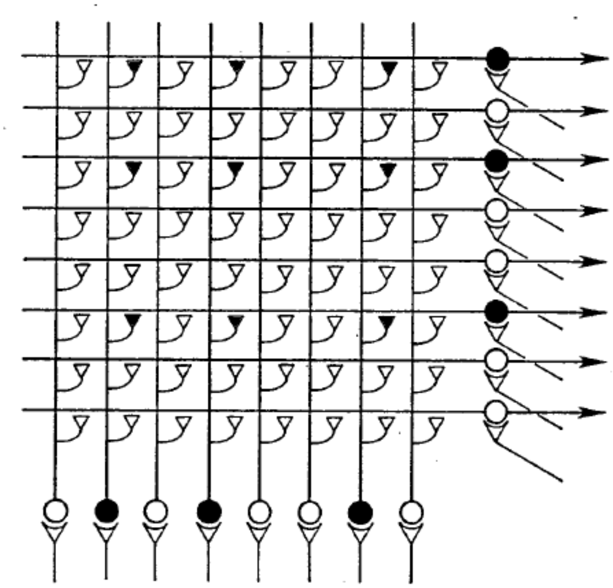

Figure 3

A simplified neural network that can learn to associate one kind of information about an object (such as the sound of the word designating its name) with another kind of information (its visual appearance). Each kind of information is represented as a pattern of activation (active units are shown here in black, inactive units in white), and the learned association depends on strengthened connections from the units active in one of the representation to the units active in the other. From McClelland, J. L. & Rumelhart, D. E. (1988). *Explorations in Parallel Distributed Processing: A Handbook of Models, Programs, and Exercises,* 90. Permission Pending.

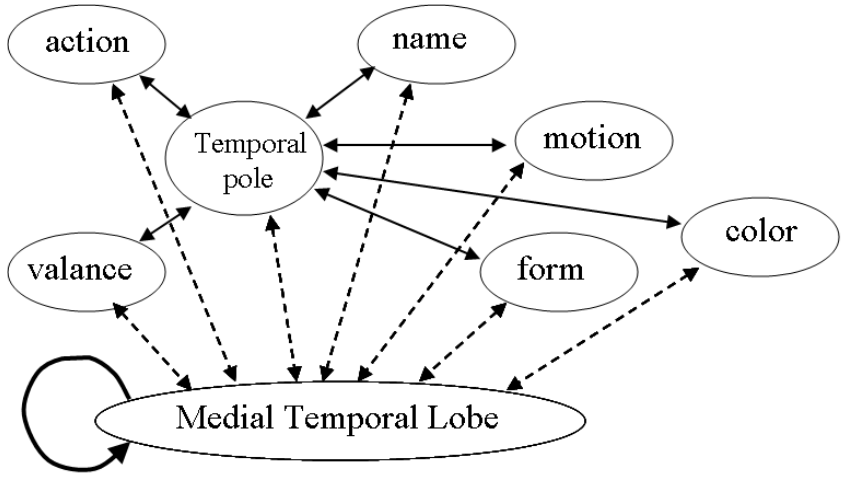

Figure 4

A schematic diagram of the brain network thought to underlie memory in the complementary learning systems theory. In the diagram, ovals represent brain areas containing many neurons, and arrows between areas represent projections or bundles of axons that carry signals between brain areas. All of the ovals other than the oval labeled Medial Temporal Lobe are thought to be parts of the slow-learning, neocortical system.