CHAPTER 27

The Neural Basis of Consciousness and Explicit Memory: Reflections on Kihlstrom, Mandler, and Rumelhart

James L. McClelland
Carnegie Mellon University

These reflections extend the approach taken in Rumelhart's chapter (25, this volume) on emotion to the nature of consciousness and explicit memory. In this chapter I argue that a cognitive neuroscience perspective provides a framework in which we can account for the principal aspects of consciousness and explicit memory stressed by Kihlstrom (chapter 24, this volume), but that allows us to consider the question of the centrality of the concept of self in a somewhat different light. I then consider Mandler's (chapter 26, this volume) search for the functions of consciousness, and suggest how it may be useful to think of consciousness, not so much as a separate faculty with its own functions, but as a manifestation of certain properties of overall system function. Finally, I comment on emotion and its relation to consciousness and memory, relating the central theme of Rumelhart's chapter to issues raised both by Kihlstrom and Mandler.

A BRAIN SYSTEMS MODEL OF INFORMATION PROCESSING AND MEMORY

A brain systems model of information processing and memory (McClelland, McNaughton, & O'Reilly, 1995) has been developed in the context of the conceptual sketch depicted in Fig. 27.1. According to this model, human cognitive processes take place within a highly interconnected neural information-processing system consisting of large numbers of simple processing elements

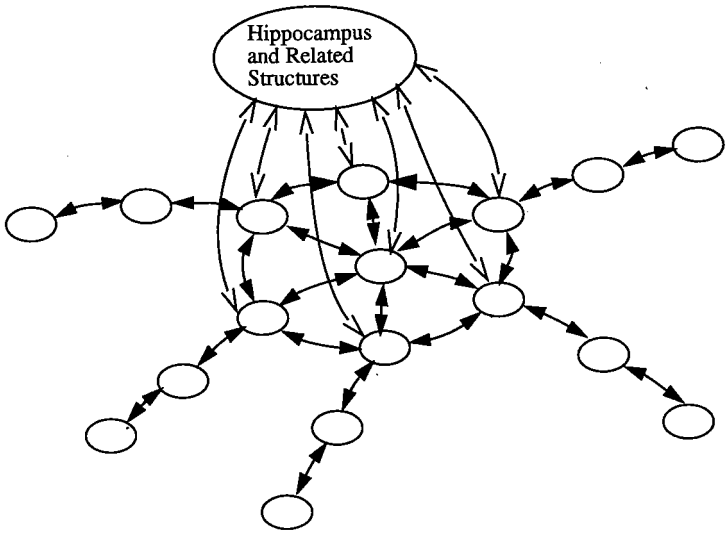

FIG. 27.1. A sketch of the brain systems model of information processing and memory proposed by McClelland et al. Adapted from McClelland, McNaughton, & O'Reilly (1994).

(neurons) organized into modules. Figure 27.1 bears some resemblance to Rumelhart's Fig. 25.1, though this figure stresses the interconnectedness of central parts of the system, and the existence of relatively separate peripheral pathways associated with different senses and effector systems. The figure also highlights the hippocampal system, which plays a crucial role in the formation and retrieval of recent explicit memories. In referring to the parts of this system, everything other than the hippocampal system is referred to as "the processing system," with the phrase "the central parts of the processing system" designating those regions that are heavily interconnected with each other. In the brain, processing system consists of most of the neocortex and several other structures that participate in information processing; the central parts of the processing system are primarily those areas of temporal, parietal, and frontal cortex often called *secondary* or *tertiary association* areas. The hippocampal system consists of the hippocampus itself and adjacent neocortical areas in the medial temporal lobes.

In this model, the presentation of a stimulus for processing results in a pattern of activation distributed widely throughout the processing system. The pattern of activation depends on the prior state (pattern of activation) in the system, the input, and the strengths of the connections among the processing units. The connections among the units (within and between modules) impose coherence on these patterns of activation; that is, there is a tendency for the patterns of activation in one part of the system to depend on the patterns

of activation in the other parts. Think of a state of mind as being this very pattern of activation itself.

For our purpose of discussing the nature and functions of consciousness, specific reference must be made to the role of the frontal lobes. The view adopted here, articulated in Cohen and Servan-Schreiber (1992), is that the frontal lobes may play a crucial role in maintaining representations of task-relevant context in a form suitable for orchestrating activity in the rest of the system. One possibility is that the frontal lobes contain modules specialized for the maintenance of activation of neurons that represent specific task-relevant information. Cohen and Servan-Schreiber introduced the phrase *memory for context* as a shorthand characterization of the role of these representations. This fits with the fact that frontal lobe damage leads to deficits in the modulation of behavior by task instructions. Inability to inhibit prepotent responses can be seen as a special case of such a deficit.

For purposes of discussing explicit and implicit memory, reference must be made to mechanisms that allow activity in the system at one point in time to affect the system's behavior at later times. In the model, the connections among the units are adaptive—that is, they are subject to modification as a result of activity in the system. For our purposes we just make use of a simple Hebbian conception of synaptic plasticity: When two units are active in temporal synchrony or close succession, the strength of the connection(s) between them is increased (Hebb, 1949). Such changes in connection strength underlie both implicit memory and explicit memory.

Implicit memory refers to cases in which prior experience influences later processing without conscious or deliberate recollection of the experience (D. L. Schacter, 1987). This can occur in our model in several ways. One such is the strengthening of connections among the units within the neocortical system (McClelland & Rumelhart, 1985). For example, an individual will achieve the same perception more readily the second time a stimulus is shown, because the first showing strengthens the connections among the units in the cognitive system whose activation constitutes the percept.

Explicit memory refers to cases in which prior experience is deliberately or consciously accessed (D. L. Schacter, 1987). In the model, explicit memory amounts to the construction of a weaker version of at least some parts of the pattern of activation that was present at the time of the initial experience.

A crucial question is, what makes an experiencer think that some mental state is an explicit memory? It cannot be actual prior occurrence, of course, because false memories often occur. One answer is that we think it is a memory to the extent that it carries with it material referring to a specific context. For example, I think that I remember the proposition "Hot amethysts are yellow" because when I remember it I also remember John Kihlstrom saying this in the course of a talk I once heard him give on source amnesia. It appears, from the effects of frontal lesions on source amnesia, that the frontal lobes

may indeed contain those parts of the brain necessary for the representation of situational context, consistent with the claims of Cohen & Servan-Schreiber (1992).

As already noted, the model assumes that connection weight changes subserve explicit memory as well as implicit memory. However, the changes made within the processing system are not sufficient to subserve the formation of a novel, arbitrary association all at once. Rather, the initial formation of such arbitrary associations is thought to depend on the hippocampal system. On this view, lesions to the hippocampal system produce such profound deficits in formation of new explicit memory because explicit memory generally involves arbitrary associations. A crucial piece of evidence for our claim that it is the formation of novel arbitrary associations that depends on an intact hippocampal system, is the fact that normal subjects show implicit learning of novel associations, but profound amnesics do not (see McClelland et al, 1995, for discussion).

The model described here assumes that the involvement of the hippocampal system in learning and memory takes the following form: At the time of an experience, changes to connections occur both within the processing system itself and in the hippocampal system. Thus, when thinking about the word pair "Locomotive–Dishtowel" in an experimental psychology experiment, the pair, together with any mental image formed of them interacting, together with any context present in the patterns of activation in the processing system during study, gives rise to a distributed pattern of activation distributed widely throughout the processing system. Connections from the processing system to the hippocampal system produce a reduced description of the pattern in the processing system in the hippocampus, and synaptic plasticity in the hippocampus associates the elements of the reduced description with each other. Later, at the time of test, a reminder of the study session and the first word of the pair are presented as retrieval cues. These produce a pattern that overlaps with the pattern that was present in the processing system during study. The connections from the processing system to the hippocampal system then project this pattern into the hippocampus, where the associative learning that took place during study leads to pattern completion. The return connections from the hippocampal system to the processing system then reinstate enough of the rest of the neocortical pattern to serve as the basis of recall of the second word of the pair.

The final point of this theory is *consolidation.* Memories that were initially dependent on the intact hippocampal system lose this dependence with time, and on the model described here this occurs through the gradual accumulation of connection changes over repeated reinstatements of overlapping patterns of activation containing the content being consolidated. Novel, arbitrary associations can be acquired within the processing system itself gradually, and thus ultimately they may come to lose their dependence on the hippocampal

system.

CONSCIOUSNESS, MEMORY, AND EMOTION IN THE CONTEXT OF THE MODEL

With the key elements of this theory (and that of Cohen & Servan-Schreiber, 1992) in mind, consider some of the issues raised by Kihlstrom and Mandler in their discussions of the nature and functions of consciousness and explicit memory.

Consciousness, Memory, and the Self

Kihlstrom (chapter 24, this volume) adopts the view, previously proposed by James, that consciousness and explicit memory are invariably associated with the sense of self. He argues that the self is a complex cognitive structure containing many versions of itself, that these are usually associated with each other but not always, that consciousness involves the self as a participant in the conscious experience, and that explicit memory involves a recollection of the role of the self in the prior experience. He suggests that implicit memory involves access to and use of memory disembodied from its connection with self. Although there are many appealing aspects of this construal, I am inclined to think that it is not quite correct.

It is not clear to me that consciousness is always associated with the self. A conscious state of mind can explicitly refer to the experiencer as a participant, as in the state of mind I have when I contemplate a glorious sunset and contemplate how glad it makes me feel to be alive; or it might not, as in the state of mind when I contemplate the same sunset and think about why the sky changes colors so when the sun is near the horizon. A great deal of what I know is associated with myself, but there is a considerable amount that is not. Similarly, there are explicit memories that do not seem to involve the self in any real way. For example, I remember a scene from a movie starring Marlon Brando as a good-cowboy-gone-bad. As he is dying, the girl (she's Spanish) tells him she will love him forever and presses into his hands the necklace he has been trying to steal from her throughout the movie. I remember him saying to her with his dying breath "You don't know how good that makes me feel." The event occurred at night in rocky terrain; he died in her arms.[1] It is fair to say that what I have just described is an explicit memory—it is a con-

[1]I make no claim as to the veracity of any aspect of this story; however, I would bet there is some such scene in some movie starring Marlon Brando, and that I saw it, perhaps 20 years ago.

scious and deliberate recollection of past experience—yet it does not involve me as a participant.

Instead of Kihlstrom's proposal, I consider the somewhat weaker view, that many but not all of our explicit memories do involve ourselves as participants. A thought about one's self may be just one of many possible sorts of context that might co-occur with some conscious content. Indeed, we might extend the same line of thinking to explicit memory: If what is stored in memory is an auto-association of whatever was present in our consciousness at the time of the experience, then if the conscious experience made reference to the self, the memory, when retrieved, may do so as well. This modification of Kihlstrom's proposal fits well with the brain systems model already described. That model does not make special reference to a place for the self, yet a representation of the self as a participant in an event, or as part of the context in which an event occurred, may well be a part of many memories.

Some of the phenomena Kihlstrom reviews are very striking and are certainly consistent that the self plays a role in consciousness and memory. For example, in his discussion of multiple personality disorders, Kihlstrom points out that many times the knowledge about one personality may not be accessible to another. We can account for such (rare) events in our brain systems model by assuming that the self is part of the representation of the inaccessible knowledge. Assume for the moment that there is a module in the processing system somewhere that is the module in which the self is represented; and that in multiple personality disorders, the patterns of activation that represent the alternative selves are mutually incompatible; each is a strong attractor state very different from the other so that only one of them can be actively represented at a time. Then when the pattern representing (for example) the personality A. J. Brown is present, this pattern will tend to serve as a context-retrieval cue for events involving A. J. Brown; but will serve as a highly inappropriate cue for events involving Ansel Bourne. Indeed, the representation of Brown, if it is strongly enough maintained, might serve to prevent the activation of the alternative representation of Ansel Bourne, If so, one personality would be inaccessible when the other is in place.

The previous account is consistent with the possibility that some explicit memories, formed when one personality is in place, may be accessible when the other is in place; we would expect this to be true especially for those memories that have no strong link to the self. In this context, Kihlstrom's statement that "A. J. Brown once gave testimony in church that referred to an incident that had actually happened to Ansel Bourne" is intriguing. The account predicts that it would be more likely for Brown to testify to an event that was merely observed by Bourne, than one that had actually happened to Bourne. So while consulting a copy of James in search of further details, I found that the text was not much more detailed, but it differed from Kihlstrom's restatement in exactly the way that fits best with the account here: "Once at a prayer

meeting he made what was considered by the hearers a good address, in the course of which he related an incident which he had *witnessed* in his natural state as Bourne" (James, 1890, Vol. 1, p. 392, italics added).

Functions of Consciousness

Turning now to a consideration of Mandler's chapter 26 (this volume), I found a strong contrast between his view of consciousness and my own. He treats consciousness as though it were a specific faculty, like, say, olfaction, and considers what its functions might be. To me, consciousness accompanies certain types of brain states and these states have certain characteristics and certain effects. Perhaps a discussion of Mandler's ideas about the selective function of consciousness and of the feedback function of consciousness will be helpful. He argues that consciousness selects partially activated, preconscious material that fits with current demands and intentions, and that the result of this selection process is that the selected material becomes primed—so that the same material now becomes more available to consciousness on later processing. Consider this alternative: due to the interconnectedness of the higher levels of the processing system, as illustrated in Fig. 27.1, the brain has a tendency to impose coherence on its states of activation. Rumelhart makes essentially this point in his chapter. Partially activated material that hangs together with other active material becomes a part of the coherent state; the material that does not hang together with the other material is suppressed, and lost from the state. Consciousness reflects the contents of these coherent states; also, material that is present in coherent states is active longer and stronger than material that is suppressed, and the amount of synaptic modification, and hence the amount of priming is increased.

For readers who are not very familiar with connectionist networks, these ideas may seem unfamiliar and even implausible. Perhaps they can be made more approachable in the context of the interactive activation model of letter perception (McClelland & Rumelhart, 1981). One can view the interactive activation model as a system of many interconnected modules that settle in a mutually interdependent fashion into a coherent state. When an ambiguous letter (e.g., a letter that is partially obscured so it might be an uppercase *R* or *K*) is presented in the fourth position of a display containing the letters WOR in positions 1–3, units representing *R* and *K* initially receive equal activation based on the input to the fourth position (see Fig. 27.2). But the entire pattern over all four positions tends to activate the word *WORK* more strongly than any other word, and *WORK* feeds back activation to the fourth position *K;* the *K* thus becomes more active than the alternative *R,* and ultimately the *R* is suppressed. The overall pattern in which the word level represents WORK and four letter positions represent *W, O, R,* and *K* becomes relatively stable, for a period of time. The *K* would be primed (in that its connections with its fea-

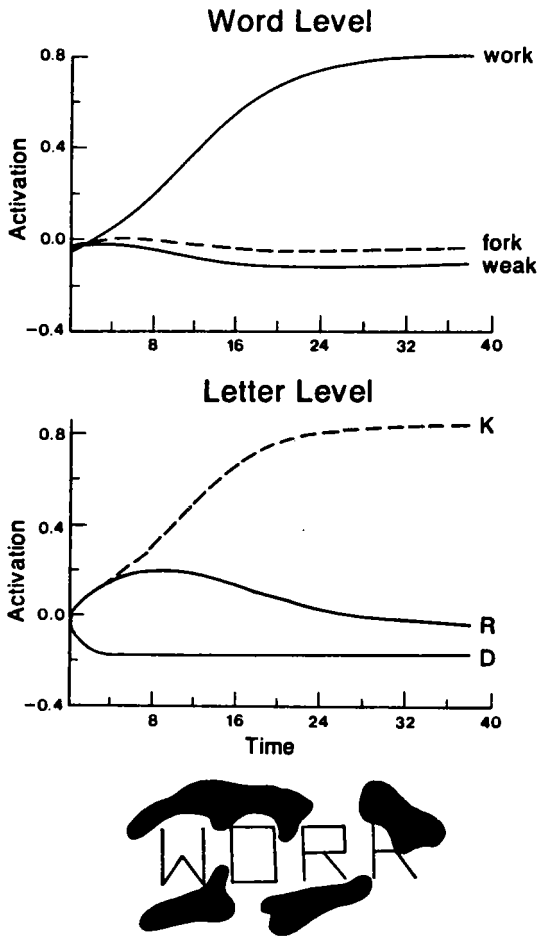

FIG. 27.2. Illustration of the settling process in the interactive activation model of visual word recognition, illustrating how coherence of activations over many modules can arise from a constraint satisfaction process. In this case, the coherent pattern of activation represents a word at the feature, letter, and word levels. The modules represent letters and words, and the coherence is maintained by connections that make each word a coherent attractor state for the network. From McClelland, Rumelhart, and Hinton (1986). Copyright 1986 by MIT Press. Reprinted with permission.

tures and with the word *WORK* would be strengthened), and the *R* would tend not to be primed very much.

These ideas fit together well with many of Mandler's other comments, and he would not disagree with the suggestions outlined here. For example, Mandler follows Marcel (1983) in treating the contents of consciousness as being constructed out of the preconscious material that is active at any given time. This is exactly what is seen in the interactive activation model when a pseudoword is shown; many words are partially activated, and each contributes to the construction of a perception of a nonword. Furthermore, in discussing the selective function of consciousness, he says that "this hypothesis of selective and limited activation of situationally relevant structures requires no homunculuslike function for consciousness in which some independent agency controls, selects, and directs thoughts and actions that have been made available to consciousness. Given an appropriate database, it should be possible to simulate this particular function of consciousness without an appeal to an independent decision-making agency." The only place where we seem to disagree is in the question of whether the functions Mandler describes should be attributed to consciousness itself or to the brain state that consciousness accompanies.

A Brain Systems Approach to Emotion

Finally, let us consider Rumelhart's (Chapter 25, this volume) brain systems approach to emotion. In brief, he suggests that emotions correspond to patterns of activation over a set of neuromodulatory systems. Each connection weight actually consists of several modulator-specific subweights, the effective strength of which depends on the intrinsic value of the weight times the concentration of the neuromodulator. The total effect of the connection weight is the sum of the effective strengths of all of its component subweights. This idea has considerable appeal because it suggests how state-dependency might be embodied easily in a single system. In each emotional (neuromodulatory) state, a different set of subweights will have the strongest effect; yet there will be some sharing across states, to the extent that each neuromodulatory system remains partially active. Knowledge that was acquired in the subweights associated with a particular modulatory state would be most readily accessible in the same state, less so in other states.

Although this idea has considerable appeal, it is worth noting that there is another way of thinking about emotion that can account for the same state dependency. This is the idea that emotional states act as contexts for other activity in the cognitive system. If an emotion is represented as a pattern of activation over a part of the system, and if this part is connected with other parts, then patterns over these other parts that were active together with a particular emotion will tend to become associated with that emotion, and will thus tend to be more easily activated when the same emotion is in place. Further,

patterns that have co-occurred with a particular emotion will tend to cause that emotion to become active; in general, the system will tend to maintain coherence between the emotion and other aspects of the overall mental state. This view of emotion is appealing for several reasons. First, as we have known since the seminal work of S. Schachter and Singer (1962), emotion is not simply a by-product of hormonal/neuromodulatory state. Several very different emotional states can arise from the same hormonal manipulation, depending on other inputs, such as knowledge that the hormonal state was produced by an injection, or environmental inputs that might tend to induce anger or happiness; this is consistent with Mandler's suggestion that emotions are constructions that combine physiological and cognitive components. Second, this view of emotional state dependence links it with other forms of context dependence, and with Kihlstrom's ideas of variants of the self. In particular, my own sense is that some emotions are more consistent with some variants of the self than others. Lastly, it should be noted that this view of emotion is not incompatible with Rumelhart's proposal; it seems likely that both ideas are part of the story.

SUMMARY AND CONCLUSIONS

In this chapter, I suggest how aspects of consciousness, memory, and emotion can be understood within the context of a connectionist/brain systems account of the organization of the cognitive system. Although a connectionist/brain systems perspective clearly leaves a lot of room for a range of views at this point, the approach appears to hold considerable promise of providing a framework in which the study of neuroanatomy and neurophysiology, and of the effects of brain lesions can be brought together with the study of cognitive processes, and even with the study of consciousness and emotion. Such a perspective might well lead to a deeper understanding of the nature and contents of conscious thought and emotion, as well as a deeper understanding of their physical basis in the brain.

REFERENCES

Cohen, J. D., & Servan-Schreiber, D. (1992). Context, cortex, and dopamine: A connectionist approach to behavior and biology in schizophrenia. *Psychological Review, 99,* 45–77.

Hebb, D. O. (1949). *The organization of behavior.* New York: Wiley.

James, W. (1890). *Psychology (briefer course).* New York: Holt.

Marcel, A. (1983). Conscious and unconscious perception: Experiments on visual masking and word recognition. *Cognitive Psychology, 15,* 197–237.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1994). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures

of connectionist models of learning and memory. (Tech. Rep. PDP.CNS. 94.1), Pittsburgh, PA: Carnegie Mellon University.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102,* 419–457.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review, 88,* 375–407.

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General, 114,* 159–188.

McClelland, J. L., Rumelhart, D. E., & Hinton, G. E. (1986). The appeal of parallel distributed processing. In D. E. Rumelhart, J. L. McClelland, & the PDP research group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 1: Foundations* (pp. 3–44). Cambridge, MA: MIT Press.

Schachter, S., & Singer, J. (1962). Cognitive, social and physiological determinants of emotional state. *Psychological Review, 69,* 379–399.

Schacter, D. L. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 13,* 501–518.