The Importance of Modeling for the Future of Molecular Studies of Learning and Memory

Randall C. O'Reilly
Department of Psychology
University of Colorado Boulder
Campus Box 345
Boulder, CO 80309
oreilly@psych.colorado.edu

James L. McClelland
Center for the Neural Basis of Cognition
and Department of Psychology
Carnegie Mellon University
Pittsburgh, PA 15213
mcclelland@cnbc.cmu.edu

20th June 2000

Draft Chapter for Alcino J. Silva Book on Molecular Studies of Learning and Memory, MIT Press

Abstract

Computational models are important for guiding and interpreting molecular studies of learning and memory because they provide a bridge between biological and behavioral levels of analysis. These models facilitate the identification of central underlying principles that span different levels, and they can accommodate the many complexities of each of these levels and their interrelationships. We present an overview of our computational approach towards understanding the different contributions of the neocortex and hippocampus in learning and memory. The approach is based on a set of principles derived from converging biological, psychological, and computational constraints. This framework provides insight into: (a) which behavioral paradigms are most appropriate for dissociating cortical and hippocampal learning contributions; (b) effects of selective impairments on hippocampal areas; (c) effects of synaptic modification (LTP/LTD) impairments on various pathways within the hippocampal system; (d) effects of manipulations of hippocampal activation parameters; and (e) effects of manipulations that selectively impair different components of learning (Hebbian versus error-driven). We hope that these insights provide fruitful avenues for further research using molecular techniques to inform our understanding of exactly how learning and memory phenomena emerge from their biological basis.

Introduction

Computational models can provide an important tool for linking data across multiple levels of analysis. The cognitive implications of findings about the cellular and molecular properties of neurons are often not immediately apparent — there is simply too much complexity for verbal theories to accommodate. The situation is analogous to having a human try to predict the weather by looking at a number of satellite measurements — there is simply too much information at many different levels, which needs to be integrated in complex ways to make sense of the phenomena. A computational model, of the weather or of the brain, can help by providing a way to explore the emergent properties of a complex system. Models must, of necessity, abstract away from some of the underlying complexity. By formulating specific grounds for the abstraction — that is, by articulating a set of principles that guide the construction of the model — a good model becomes a vehicle for exploring the adequacy of a particular view of the fundamental nature of the complex system that depends upon all the underlying details. Once a model is formulated, is can provide an explicit, formal tool for exploring how behavior arises from the molecular and neurophysiological substrate, and for understanding why manipulations of underlying mechanisms (e.g., genetic knockouts or lesions) give rise to the observed behavioral effects.

Although there is clearly great promise for computational models, they face a number of serious obstacles. Perhaps most importantly, models are only as good as the validity of the principles they embody ("garbage-in, garbage-out"). In addition, it is essential that the operation of the model be intelligible, so we can understand why it behaves as it does. To deal with these problems, it is essential to develop a detailed understanding of the model that goes beyond a mere reporting of the results of particular simulations.

This paper presents a computational approach towards understanding the different contributions of the neocortex and hippocampus in learning and memory. This approach uses basic principles of computational neural network learning mechanisms to understand both *what* is different about the way these two neural systems learn, and *why* they should have these differences. Thus, the computational approach can go beyond mere description towards understanding the deeper principles underlying the organization of the cognitive system. These principles are based on an convergence of biological, psychological, and computational constraints, and serve to bridge between these different levels of analysis. We then suggest how this computational framework can be used to make sense of molecular and cellular manipulations to these brain areas.

The set of principles discussed in this paper was first developed in O'Reilly and McClelland (1994) and McClelland, McNaughton, and O'Reilly (1995), and they have been refined and elaborated in several further publications (McClelland & Goddard, 1996; O'Reilly, Norman, & McClelland, 1998; Hasselmo & McClelland, 1999; O'Reilly & Rudy, in press, 2000). The computational principles have been applied to a wide range of learning and memory phenomena across several species (rats, monkeys and humans). For example, they can account for impaired and preserved learning capacities with hippocampal lesions in conditioning, habituation, contextual learning, recognition memory, recall, and retrograde amnesia.

The Principles

There are several levels of principles that can be distinguished by their degree of specificity in characterizing the nature of the underlying mechanisms. We begin with the most basic principles and proceed towards greater specificity.

Learning Rate, Overlap, and Interference

The most basic set of principles can be motivated by considering how subsequent learning can interfere with prior learning. A classic example of this kind of interference can be found in the $AB - AC$ associative learning task (e.g., Barnes & Underwood, 1959). The A represents one set of words that are associated with two different sets of other words, B and C. For example, the word *window* will be associated with the word *reason* in the AB list, and associated with *locomotive* on the AC list. After studying the AB list of associates, subjects are tested by asking them to give the appropriate B associate for each of the A words. Then, subjects study the AC list (often over multiple iterations), and are subsequently tested on both lists for recall of the associates after each iteration of learning the AC list. Subjects exhibit some level of interference on the initially learned AB associations as a result of learning the AC list, but they still remember a reasonable percentage (see Figure 1a for representative data).

The first set of principles concern the effects of overlapping representations (i.e., shared units between two different distributed representations) and rate of learning on the ability to rapidly learn new information with a level of interference characteristic of human subjects:

- Overlapping representations lead to interference (conversely, separated representations prevent interference).

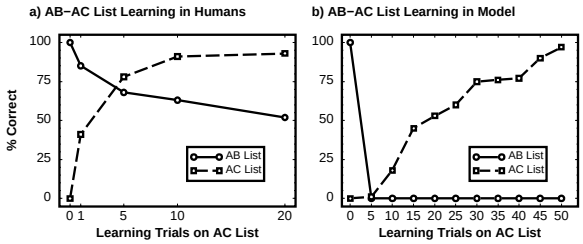

Figure 1: Human and model data for AB-AC list learning. a) Humans show some interference for the AB list items as a function of new learning on the AC list items. b) Model shows a catastrophic level of interference. (data reproduced from McCloskey & Cohen, 1989).

- A faster learning rate causes more interference (conversely, a slower learning rate causes less interference).

The mechanistic basis for these principles within a neural network perspective is straightforward. Interference is caused when weights used to encode one association are disturbed by the encoding of another (Figure 2a). Overlapping patterns share more weights, and therefore lead to greater amounts of interference. Clearly, if entirely separate representations are used to encode two different associations, then there will be no interference whatsoever (Figure 2b). The story with learning rate is similarly straightforward. Faster learning rates lead to more weight change, and thus greater interference (Figure 3). However, a fast learning rate is necessary for rapid learning.

Integration and Extracting Statistical Structure

Figure 3 shows the flip side of the interference story, *integration*. If the learning rate is low, then the weights will integrate over many experiences, reflecting the *underlying statistics* of the environment (White, 1989; McClelland et al., 1995). Furthermore, overlapping representations facilitate this integration process, because the same weights need to be reused across many different experiences to enable the integration produced by a slow learning rate. This leads to the next principle:

- Integration across experiences to extract underlying statistical structure requires a slow learning rate and overlapping representations.

Episodic Memory and Generalization: Incompatible Functions

Thus, focusing only on pattern overlap for the moment, we can see that networks can be optimized for two

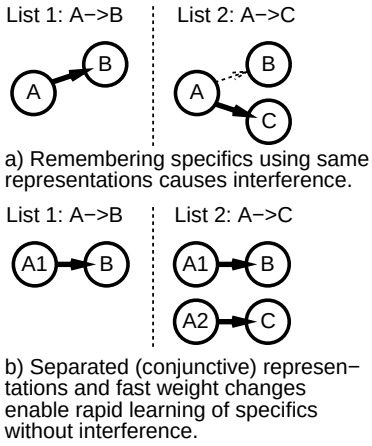

Figure 2: Interference as a function of overlapping (same) representations versus separated representations. **a)** Using the same representation to encode two different associations ($A \rightarrow B$ and $A \rightarrow C$) causes interference — the subsequent learning of $A \rightarrow C$ interferes with the prior learning of $A \rightarrow B$ because the A stimulus must have stronger weights to C than to B for the second association, as is reflected in the weights. **b)** A separated representation, where A is encoded separated for the first list ($A1$) versus the second list ($A2$) prevents interference.

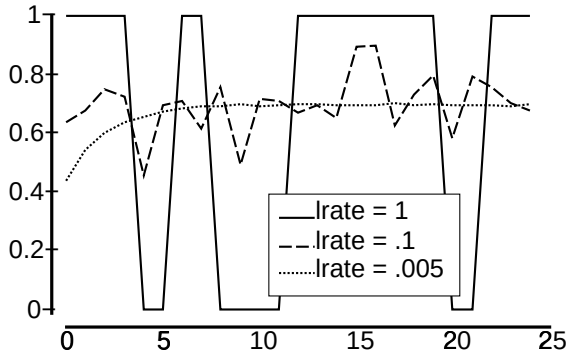

Figure 3: Weight value learning about a single input unit that is either active or not. The weight increases when the input is on, and decrease when it is off, in proportion to the size of the learning rate. The input has an overall probability of being active of .7. Larger learning rates (.1 or 1) lead to more interference on prior learning, resulting in a weight value that bounces around substantially with each training example. In the extreme case of a learning rate of 1, the weight only reflects what happened on the previous trial, retaining no memory for prior events at all. As the learning rate gets smaller (.005), the weight smoothly averages over individual events and reflects the overall statistical probability of the input being active.

different, and incompatible, functions: avoiding interference or integrating across experiences to extract generalities. Avoiding interference requires separated representations, while integration requires overlapping representations. These two functions each have clear functional advantages, leading to a further set of principles:

- Interference avoidance is essential for *episodic* memory, which requires learning about the specifics of individual events and keeping them separate from other events.
- Integration is essential for encoding the general statistical structure of the environment, abstracted away from the specifics of individual events, which enables *generalization* to novel situations.

The incompatibility between these functions is further evident in these descriptions (i.e., encoding specifics versus abstracting away from them). Also, episodic memory requires relatively rapid learning — an event must be encoded as it happens, and does not typically repeat itself for further learning opportunities. This completes a pattern of opposition between these functions: episodic learning requires rapid learning while integration and generalization requires slow learning. This is summarized in the following principle:

- Episodic memory and extracting generalities are in opposition. Episodic memory requires rapid learning and separated patterns, while extracting generalities requires slow learning and overlapping patterns.

The Hippocampus and Neocortex

Armed with these principles, the finding that neural network models that have highly overlapping representations exhibit *catastrophic* levels of interference (McCloskey & Cohen, 1989, Figure 1b) should not be surprising. A number of researchers showed that this interference can be reduced by introducing various factors that result in less pattern overlap (e.g., Kortge, 1993; French, 1992; Sloman & Rumelhart, 1992; McRae & Hetherington, 1993). Thus, instead of concluding that neural networks are fundamentally flawed in some way, McClelland et al. (1995) argued that this catastrophic failure serves as an important clue into the structure of the human brain.

Specifically, we argued that because of the fundamental incompatibility between episodic memory and extracting generalities, the brain should employ two separate systems that each optimize these two objectives individually, instead of having a single system that tries to strike an inferior compromise. This line of reasoning provides a good fit to the known properties of the hippocampus and neocortex, respectively. The details of this fit in various contexts is summarized briefly in what follows, but the general idea is that:

- The hippocampus rapidly binds together information using pattern-separated representations to minimize interference.
- The neocortex slowly learns about the general statistical structure of the environment using overlapping distributed representations.

(see also Sherry & Schacter, 1987 for a similar conclusion).

This view of hippocampal function is consistent with the *conjunctive* or *configural* representations theory (Sutherland & Rudy, 1989; Rudy & Sutherland, 1995; Wickelgren, 1979; O'Reilly & Rudy, in press). A conjunctive/configural representation is one that binds together (conjoins or configures) multiple elements into a novel unitary representation. We have shown that pattern separation and conjunctive representations are two sides of the same coin, and that both are caused by the use of *sparse* representations (having relatively few active neurons) that are a known property of the hippocampus (O'Reilly & McClelland, 1994; O'Reilly & Rudy, in press). To summarize:

- Sparse hippocampal representations lead to pattern separation (to avoid interference) and conjunctive representations (to bind together features into a unitary representation).

Pattern completion is an additional principle that is required for recalling information from conjunctive hippocampal representations (McNaughton & Morris, 1987; Rolls, 1989; Treves & Rolls, 1994), yet it conflicts with the process of pattern separation that forms these representations in the first place (O'Reilly & McClelland, 1994). Pattern completion occurs when a partial input cue drives the hippocampus to complete to an entire previously-encoded set of features that were bound together in a conjunctive representation. For a given input pattern, a decision must be made to recognize it as a retrieval cue for a previous memory and perform pattern completion, or to perform pattern separation and store the input as a new memory. This decision is often difficult given noisy inputs and degraded memories. The hippocampus implements this decision as the effects of a set of basic mechanisms operating on input patterns (O'Reilly & McClelland, 1994; Hasselmo & Wyble, 1997), and it does not always do what would

seem to be the right thing to do from an omniscient perspective knowing all the relevant task factors — this can complicate the involvement of the hippocampus in various problems.

Learning Mechanisms: Hebbian and Error Driven

To more fully explain the roles of the hippocampus and neocortex we need to understand how learning works in these systems. Among the basic classes of learning mechanisms that have been discussed in the literature, two of the most prominent are Hebbian and error-driven learning (e.g., Marr, 1971; McNaughton & Morris, 1987; Gluck & Myers, 1993; Schmajuk & DiCarlo, 1992). Briefly, Hebbian learning (Hebb, 1949) works by increasing weights between co-active neurons (and usually decreasing weights when a receiver is active and the sender is not), which is a well-established property of biological synaptic modification mechanisms (e.g., Collingridge & Bliss, 1987). Hebbian learning is useful for binding together features active at the same time (e.g., within the same episode), and has therefore been widely suggested as a hippocampal learning mechanism (e.g., Marr, 1971; McNaughton & Morris, 1987).

Error-driven learning works by adjusting weights to minimize the errors in a network's performance. Error-driven learning is sensitive to task demands in a way that Hebbian learning is not, and this makes it a much more capable form of learning for actually achieving some desired output for given inputs. Thus, it may be natural to associate this form of learning with the kind of procedural or task-driven learning that the neocortex is often thought to specialize in. The best example of an algorithm of this type is the *error backpropagation* algorithm (Rumelhart, Hinton, & Williams, 1986), in which the discrepancy between the desired output and the output that is actually produced is computed, and is then back-propagated through the network, allowing the calculation of how much and in what direction each weight in the network should be changed to move the entire set of weights in the direction that reduces the error the most quickly. Although the backpropagation mechanism has been widely challenged as biologically implausible (e.g., Crick, 1989; Zipser & Andersen, 1988), there are a number of other means to allow outcome information to affect the adjustment of connection weights. For example, O'Reilly (1996) showed that recurrent activation propagation in a bidirectionally-connected network communicates backpropagation error gradients in a local, biologically-plausible fashion. Alternatively, outcome information can be used to modulate Hebbian learning in a variety of ways that can lead effective learning. One method of this type is known as reinforcement learning (e.g., Mazzoni, Andersen, & Jordan, 1991).

Although the association of Hebbian learning with the hippocampus and error-driven learning with the cortex is appealing in some ways, it turns out that both kinds of learning can play important roles in both systems (O'Reilly & Rudy, in press; O'Reilly & Munakata, 2000; O'Reilly, 1998). Thus, the specific learning principles adopted here are that both forms of learning operate in both systems:

- Hebbian learning binds together co-occurring features (in the hippocampus) and generally learns about the co-occurrence statistics in the environment across many different patterns (in neocortex).
- Error-driven learning shapes learning according to specific task demands (shifting the balance of pattern separation and completion in the hippocampus, and developing task-appropriate representations in the neocortex).

A Summary of Principles

The above principles can be summarized with the following three general statements of neocortical and hippocampal learning properties (O'Reilly & Rudy, in press):

Learning rate. The cortical system typically learns slowly, while the hippocampal system typically learns rapidly.

Componential vs conjunctive representation. The cortical system has a bias towards integrating over specific instances to extract generalities but using representations in which particular components of the input are represented similarly whenever they occur. The hippocampal system is biased by its intrinsic sparseness to develop conjunctive representations of of environmental inputs in which the individual elements are not independently represented. However, this conjunctive bias trades-off with the countervailing process of pattern completion, so the hippocampus does not always develop new conjunctive representations (sometimes it completes to existing ones).

Learning mechanisms. Both cortex and hippocampus use error-driven and Hebbian learning. The error-driven aspect responds to task demands, and will cause the network to learn to represent whatever is needed to achieve goals or ends. Thus, the cortex can overcome its bias and develop specific, conjunctive representations if the task demands require this. Also, error-driven learning can shift the hippocampus from performing pattern separation to performing pattern completion, or vice-versa, as dictated by

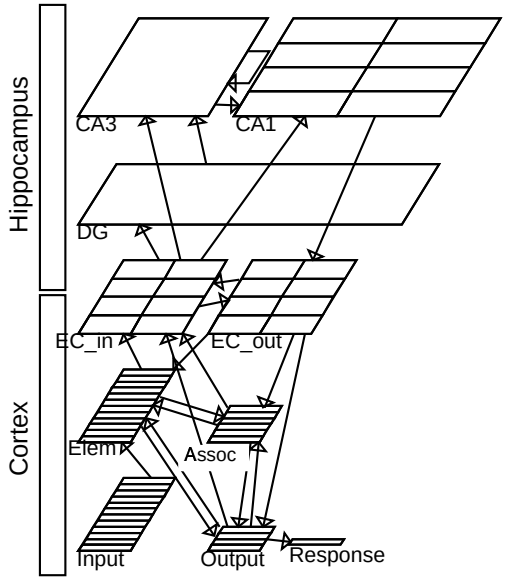

Figure 4: The O'Reilly and Rudy (in press) model, showing both cortical and hippocampal components. The cortex has 12 different input dimensions (sensory pathways), with 4 different values per dimension. These are represented separately in the elemental cortex (Elem). Higher level association cortex (Assoc) can form conjunctive representations of these elements, if demanded by the task. The interface to the hippocampus is via the entorhinal cortex, which contains a one-to-one mapping of the elemental, association, and output cortical representations. The hippocampus can reinstate a pattern of activity over the cortex via the EC.

the task. Hebbian learning is constantly operating, and reinforcing the representations that are activated in the two systems.

These principles are focused on distinguishing neocortex and hippocampus — we have also articulated a more complete set of principles that are largely common to both systems (McClelland, 1993; O'Reilly, 1998; O'Reilly & Munakata, 2000). Models incorporating these principles have been extensively applied to a wide range of different cortical phenomena, including perception, language, and higher-level cognition. In the next section, we highlight the application of the principles presented here to understanding how learning and memory phenomena emerge from their biological basis.

Applications of the Principles

We have developed neural network simulation models based on the above principles (figure 4), and used these models to understand a number of different phenomena in animal and human learning and memory. In animal learning, we have explored nonlinear discrimination learning, contextual fear conditioning, conjunctive habituation, and transitivity (O'Reilly & Rudy, in press). In human learning, we have explored cued-recall and dual-process (cortex and hippocampus) models of recognition memory applied to paired-associates, list length/strength effects, and stimulus similarity effects (Norman, O'Reilly, & Huber, 2000; O'Reilly et al., 1998; O'Reilly & Munakata, 2000). In all of these cases, essentially the same model was used, providing a compelling demonstration that the principles are sufficient to account for a wide range of findings.

To set the stage for subsequent discussion of the implications of these models for molecular studies, we briefly review some of the key animal learning simulations. Perhaps the most important message from these simulations is that one must be careful in selecting behavioral tasks to assess hippocampal learning function — the cortex alone can make important contributions to learning and memory, and the distinction between hippocampus and cortex is not as clear-cut as it might seem. Then, the subsequent section provides a range of suggestions as to how these models could inform future molecular studies.

Existing Animal Learning Simulations

One of the most important contributions of the models has been to reconcile recent nonlinear discrimination learning data with the widely-held view that the hippocampus contributes to memory by binding together elements of an experience into a unitary episodic memory. Sutherland and Rudy (1989) postulated that nonlinear discrimination learning problems provide a direct test of this kind of conjunctive binding theory. These problems require conjunctive representations because each of the individual stimuli is ambiguous (equally often rewarded and not rewarded). The negative patterning problem, $A+$, $B+$, $AB-$ (where A and B represent stimuli such as a light and a tone, and $(+)$ indicates reward while $(-)$ indicates lack of reward) is a good example. This problem requires that the conjunction of the two stimuli ($AB-$) be treated differently from the two stimuli separately ($A+$, $B+$). A conjunctive representation that forms a novel encoding of the two stimuli together can facilitate this form of learning. Therefore, it follows that the hippocampus should play a critical role in these kinds of tasks. However, it is now clear that a number of nonlinear discrimination learning problems are unimpaired by hippocampal damage (Rudy & Sutherland, 1995).

The general explanation of these results according to the full set of principles outlined above is that:

- The explicit task demands present in a nonlinear discrimination learning problem cause the cortex

alone (with a lesioned hippocampus) to learn the task via error-driven learning.

- Nonlinear discrimination problems take many trials to learn even in intact animals, allowing the slow cortical learning to accumulate a solution.
- The absence of hippocampal learning speed advantages in normal rats, despite the more rapid hippocampal learning rate, can be explained by the fact that the hippocampus is engaging in pattern completion in these problems, instead of pattern separation. Pattern completion is triggered by the very high levels of stimulus overlap across training items in these problems.

The results of the simulation models support this explanation, demonstrating that the cortex alone can learn a number of nonlinear discrimination problems at the same rate as the intact model (O'Reilly & Rudy, in press). There are some interesting wrinkles to this simple story, where some nonlinear discrimination problems do show sensitivity to hippocampal damage (see O'Reilly & Rudy, in press for details), but the clear message is that tasks that seem on the surface to provide good indicators of hippocampal function are not actually very useful when the full set of computational principles outlined above is taken into account.

These principles do however suggest another set of tasks that should provide a much better measure of hippocampal learning compared to the nonlinear discrimination problems. As we just saw, the very fact that these problems *require* conjunctive representations is what drives the cortex alone to be able to solve them via error-driven learning. Therefore, O'Reilly and Rudy (in press) suggest that *incidental* conjunctive learning tasks, where conjunctive representations are not forced by specific task demands, may provide a much better index of hippocampal function. Furthermore, the task should only allow for a relatively brief period of learning, which will emphasize the rapid learning of the hippocampus as compared to the slow learning of the cortex. These tasks can thus be characterized as *rapid, incidental conjunctive learning tasks*.

There are several recent studies of tasks that fit the rapid, incidental conjunctive characterization (Save, Poucet, Foreman, & Buhot, 1992; Honey, Watt, & Good, 1998; Honey & Good, 1993; Good & Bannerman, 1997; Hall & Honey, 1990; Honey, Willis, & Hall, 1990). In some of these tasks, for example, subjects are exposed to a set of features in a particular configuration, and then the features are rearranged. Subjects are then tested to determine if they detect the rearrangement. If the test indicates that the rearrangement was detected, then one can infer the subject learned a conjunctive representation of the original configuration. The literature indicates that the incidental learning of stimulus conjunctions, unlike many nonlinear discrimination problems, *is* dependent on the hippocampus. O'Reilly and Rudy (in press) have shown that the same neural network model constructed according to our principles and tested on the nonlinear discrimination learning problems as described above exhibits a clear hippocampal sensitivity in these rapid incidental conjunctive learning tasks.

Evidence for the involvement of the hippocampal formation in the incidental learning of stimulus conjunctions has also emerged in the contextual fear conditioning literature. This example also provides a simple demonstration of the widely-discussed role of the hippocampus in spatial learning (e.g., O'Keefe & Nadel, 1978; McNaughton & Nadel, 1990). Rats with damage to the hippocampal formation do not express fear to a context or place where shock occurred, but will express fear to an explicit cue (e.g., a tone) paired with shock (Kim & Fanselow, 1992; Phillips & LeDoux, 1994; but see Maren, Aharonov, & Fanselow, 1997). Rudy and O'Reilly (1999) recently provided specific evidence that, in intact rats, the context representations are conjunctive in nature, which has been widely assumed (e.g., Fanselow, 1990; Kiernan & Westbrook, 1993; Rudy & Sutherland, 1994). For example, we compared the effects of preexposure to the conditioning context with the effects of preexposure to the separate features that made up the context. Only preexposure to the intact context facilitated contextual fear conditioning, suggesting that conjunctive representations across the context features were necessary. We also showed that pattern completion of hippocampal conjunctive representations can lead to generalized fear conditioning. Our model also appears to be compatible with recent findings by Frankland, Cestari, Filipkowski, McDonald, and Silva (1998), in which they showed that animals with hippocampal lesions were not impaired in contextual fear conditioning in cases where the context was identifiable with a simple, salient cue.

O'Reilly and Rudy (in press) have simulated the incidental learning of conjunctive context representations in fear conditioning using the same principles as described above. One important result from these models is that it is possible for the cortex alone to exhibit contextual fear effects, such that more specific tests of *conjunctive* representations of context as performed by Rudy and O'Reilly (1999) should be used to more specifically identify the contribution of the hippocampus.

Finally, we note that we have not yet applied our model to the widely-used Morris water maze task (Morris, 1984), because this task involves many complex navigational processes. It is for this same reason that we do not consider this task to be a good indicator of hip-

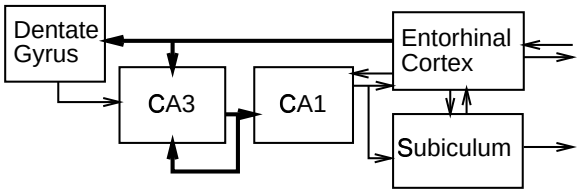

Figure 5: Schematic of the principal hippocampal areas. Adapted from Figure 5B (p. 300) of McNaughton, B. L. (1989), Neuronal Mechanisms for Spatial Computation and Information Storage. In L. Nadel, L. A. Cooper, P. Culicover, and R. M. Harnish (Eds.) *Neural Connections, Mental Computations.* Cambridge, MA: MIT Press.

pocampal function. Specifically, there are likely many redundant ways to solve the task, biological interventions can have their effects on large number of different systems relevant to task performance, and it is prone to many kinds of performance confounds.

Implications of the Models for Future Molecular Studies

The existing applications of the models reviewed above provide one key implication for future molecular studies: testing genetic or other manipulations to the hippocampus should be done using appropriate behavioral tasks that directly measure the unique contributions of the hippocampus. Here, we suggest a number of more detailed implications of the models that could be tested in future molecular studies.

Functional Contributions of Hippocampal Areas

In most biological intervention studies, the hippocampus is treated like a unitary black-box, with genetic or other manipulations designed to simply impair its overall function. In contrast, computational models make much more fine-grained predictions regarding interventions in different hippocampal areas. The structure of the hippocampal system does limit our ability to independently assess the contributions of different areas somewhat, however, because most areas are on a critical path of information flow through the system (figure 5). The only area that is not is the dentate gyrus. However, molecular methods hold out the promise of inducing synapse-specific interventions (i.e., that affect only one type of projection within the hippocampal system), which would allow more interesting questions to be addressed. We outline three specific ideas here.

First, we consider the dentate gyrus (DG). According to our models, the very sparse activations in this area produce substantial amounts of pattern separation, and not much pattern completion, making it more important for encoding new memories than retrieving existing ones (O'Reilly & McClelland, 1994). Thus, the DG can be thought of as establishing a well-separated representation in CA3, and this representation is encoded largely through synaptic changes in CA3 that enable subsequent pattern completion. Furthermore, we have suggested that the DG may not be engaged by partial cues during cued-recall (O'Reilly & McClelland, 1994). Thus, we make the following predictions:

- Complete, selective lesions of the DG should impair learning of highly similar stimuli (which depend critically on pattern separation), but not of dissimilar stimuli (which do not).

- Selective impairment of synaptic modification in the perforant path inputs to the DG, and possibly to a lesser extent the mossy fiber outputs to the CA3, should not impair learning of even highly similar stimuli. Specifically, direct neural recording of CA3 cells (e.g., place cells), or recognition-like behavioral tests where the entire original stimulus is represented, should be normal. To the extent that the DG does not participate in pattern-completion based cued-recall (e.g., presenting a partial cue at test), such tests should also be relatively unimpaired. In short, knocking out learning in the DG should have relatively little effect, as its contribution is mostly during initial encoding.

These predictions could be tested using (rapid-incidental-conjunctive) habituation studies with stimuli composed of a number of features, where similarity can be manipulated as a function of number of features in common. Cued-recall could be tested by presenting subsets of features at test.

Projection-specific techniques could also be applied to understanding the function of the CA3 and CA1. Our models posit that the CA3 is the primary site of new memory encoding, in that it encodes a novel, pattern-separated representation of an event or stimulus. This representation is encoded across a number of active neurons, which are bound together via synaptic modification within the CA3 collaterals, and synaptic changes in the perforant-path afferents facilitate subsequent recall of this encoding through pattern completion. This novel pattern-separated representation must somehow be able to reactivate corresponding cortical representations (e.g., of the component stimulus features) during recall — we think the CA1 provides a means of translating the CA3 encoding back into the language of the cortex. Thus, synaptic changes in the Schaffer collaterals connecting

the CA3 and CA1 are critical for enabling subsequent recall. Therefore, although complete damage to either the CA3 or CA1 would be devastating for the overall memory performance of the hippocampus, projection-specific knockout of synaptic modification in the interconnected pathways could have an interesting pattern of effects:

- Selective impairment of learning in the CA3 collaterals should significantly impair pattern-completion based cued-recall, while not altering the encoding properties of the CA3 (e.g., as measured by neural recording upon representation of the entire original stimulus).
- Selective impairment of learning in the Schaffer collaterals should impair all functional use of subsequently acquired hippocampal memories in the cortex, while not altering the encoding properties of CA3 (again as measured by neural recordings). A complication here would be any CA3 outputs via the fornix or subiculum — these would need to be neutralized.

Although these studies would be ambitious, and would require the use of good behavioral measures of cued-recall performance, they would tell us a great deal about how the hippocampus functions. The clear failure of any of these predictions would require rethinking of how the hippocampus functions.

Manipulations of Activation Dynamics: Sparsity

Our models depend critically on the idea that sparse activations produce pattern separation and conjunctive representations in the hippocampus. This idea could be easily tested by developing interventions that alter the overall activation level of different hippocampal areas (without causing epileptiform activity). We would predict that increasing the activation of hippocampal neurons, particularly in the DG and CA3, should much more substantially affect the ability to discriminate between similar stimuli relative to dissimilar ones. Neural network modelers have long realized that activation dynamics are as important to learning as synaptic modification mechanisms are — establishing the relevance of these activation dynamics in behavioral studies could help convey the importance of this point to a much larger audience.

Hebbian versus Error-Driven Learning Manipulations

Another category of possible molecular manipulations concerns the differential roles of Hebbian and error-driven learning mechanisms. As we mentioned earlier, there are various proposals regarding the detailed nature of biologically-based error-driven learning mechanisms, and conclusive evidence in support any of them does not yet exist. Therefore, testing for ways of isolating and further characterizing the error-driven mechanisms using molecular methods would be a promising area of future study.

For example, one specific proposal regarding the biological basis of error-driven learning could be easily tested using current LTP/LTD electrophysiology methods (O'Reilly, 1996). This proposal depends on temporal properties of intracellular calcium dynamics, to a greater degree than established Hebbian-like mechanisms. Specifically, the LTD needed to decrease weights after an error is produced by an initial elevation of calcium concentration (associated with an expected outcome) that then decreases (when that outcome is not realized, i.e., an error). Existing evidence suggests that this middling level of calcium should produce LTD, while higher levels produce LTP (e.g., Artola, Brocher, & Singer, 1990; Lisman, 1989; Bear & Malenka, 1994). Thus it is possible that error-driven and Hebbian learning mechanisms could be dissociated by selectively altering various steps in the synaptic modification cascade. One could search for such dissociations by examining the rapid-incidental-conjunctive learning tasks (which depend mostly on Hebbian learning) as compared to the nonlinear discrimination learning tasks (which depend critically on error-driven mechanisms). Establishing that these two forms of learning really exist and are dissociable, and further understanding the underlying molecular basis for such a dissociation, would constitute an important advance in our mechanistic understanding of learning and memory.

Summary

We have shown that a small set of computationally-motivated principles can account for a wide range of empirical findings regarding the differential properties of the neocortex and hippocampus in learning and memory. The detailed properties of these models make a number of specific predictions and raise a number of additional questions that could potentially be tested through molecular methods. Although many of these tests would require sophisticated projection-specific manipulations, the results would further constrain the efforts of modelers such as ourselves, leading ultimately to a deeper understanding of the mechanisms of learning and memory in the brain.

References

Artola, A., Brocher, S., & Singer, W. (1990). Different voltage-dependent thresholds for inducing long-term depression and long-term potentiation in slices of rat visual cortex. *Nature*, *347*, 69–72.

Barnes, J. M., & Underwood, B. J. (1959). Fate of first-list associations in transfer theory. *Journal of Experimental Psychology*, *58*, 97–105.

Bear, M. F., & Malenka, R. C. (1994). Synaptic plasticity: LTP and LTD. *Current Opinion in Neurobiology*, *4*, 389–399.

Collingridge, G. L., & Bliss, T. V. P. (1987). NMDA receptors — their role in long-term potentiation. *Trends in Neurosciences*, *10*, 288–293.

Crick, F. H. C. (1989). The recent excitement about neural networks. *Nature*, *337*, 129–132.

Fanselow, M. S. (1990). Factors governing one-trial contextual conditioning. *Animal Learning and Behavior*, *18*, 264–270.

Frankland, P. W., Cestari, V., Filipkowski, R. K., McDonald, R. J., & Silva, A. J. (1998). The dorsal hippocampus is essential for context discrimination but not for contextual conditioning. *Behavioral Neuroscience*, *112*, 863–874.

French, R. M. (1992). Semi-distributed representations and catastrophic forgetting in connectionist networks. *Connection Science*, *4*, 365–377.

Gluck, M. A., & Myers, C. E. (1993). Hippocampal mediation of stimulus representation: A computational theory. *Hippocampus*, *3*, 491–516.

Good, M., & Bannerman, D. (1997). Differential effects of ibotenic acid lesions of the hippocampus and blockade of n-methyl-d-aspartate receptor-dependent long-term potentiation on contextual processing in rats. *Behavioral Neuroscience*, *111*, 1171–1183.

Hall, G., & Honey, R. C. (1990). Context-specific conditioning in the conditioned-emotional-response procedure. *Journal of Experimental Psychology: Animal Behavior Processes*, *16*, 271–278.

Hasselmo, M. E., & McClelland, J. L. (1999). Neural models of memory. *Current Opinion in Neurobiology*, *9*, 184.

Hasselmo, M. E., & Wyble, B. (1997). Free recall and recognition in a network model of the hippocampus: Simulating effects of scopolamine on human memory function. *Behavioural Brain Research*, *89*, 1–34.

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

Honey, R. C., & Good, M. (1993). Selective hippocampal lesions abolish the contextual specificity of latent inhibition and conditioning. *Behavioral Neuroscience*, *107*, 23–33.

Honey, R. C., Watt, A., & Good, M. (1998). Hippocampal lesions disrupt an associative mismatch process. *Journal of Neuroscience*, *18*, 2226–2230.

Honey, R. C., Willis, A., & Hall, G. (1990). Context specificity in pigeon autoshaping. *Learning and Motivation*, *21*, 125–136.

Kiernan, M. J., & Westbrook, R. F. (1993). Effects of exposure to a to-be-shocked environment upon the rat's freezing response: Evidence for facilitation, latent inhibition, and perceptual learning. *Quarterly Journal of Psychology*, *46B*, 271–288.

Kim, J. J., & Fanselow, M. S. (1992). Modality-specific retrograde amnesia of fear. *Science*, *256*, 675–677.

Kortge, C. A. (1993). Episodic memory in connectionist networks. *Proceedings of the Twelfth Annual Conference of the Cognitive Science Society* (pp. 764–771). Hillsdale, NJ: Erlbaum.

Lisman, J. E. (1989). A mechanism for the Hebb and the anti-Hebb processes underlying learning and memory. *Proceedings of the National Academy of Sciences*, *86*, 9574–9578.

Maren, S., Aharonov, G., & Fanselow, M. S. (1997). Neurotoxic lesions of the dorsal hippocampus and Pavlovian fear conditioning. *Behavioural Brain Research*, *88*, 261–274.

Marr, D. (1971). Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society (London) B*, *262*, 23–81.

Mazzoni, P., Andersen, R. A., & Jordan, M. I. (1991). A more biologically plausible learning rule for neural networks. *Proceedings of the National Academy of Sciences*, *88*, 4433–4437.

McClelland, J. L. (1993). The GRAIN model: A framework for modeling the dynamics of information processing. In D. E. Meyer, & S. Kornblum (Eds.), *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience* (pp. 655–688). Hillsdale, NJ: Lawrence Erlbaum Associates.

McClelland, J. L., & Goddard, N. H. (1996). Considerations arising from a complementary learning systems perspective on hippocampus and neocortex. *Hippocampus*, *6*, 654–665.

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.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.), *The psychology of learning and motivation, vol. 24* (pp. 109–164). San Diego, CA: Academic Press.

McNaughton, B. L., & Morris, R. G. M. (1987). Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends in Neurosciences*, *10*(10), 408–415.

McNaughton, B. L., & Nadel, L. (1990). Hebb-Marr networks and the neurobiological representation of action in space. In M. A. Gluck, & D. E. Rumelhart (Eds.), *Neuroscience and connectionist theory* (Chap. 1, pp. 1–63). Hillsdale, NJ: Erlbaum.

McRae, K., & Hetherington, P. A. (1993). Catastrophic interference is eliminated in pretrained networks. *Proceedings of the Fifteenth Annual Conference of the Cognitive Science Society* (pp. 723–728). Hillsdale, NJ: Erlbaum.

Morris, R. G. M. (1984). Developments of a water-maze procedure for studying spatial learning in the rat. *Journal of Neuroscience Methods*, *11*, 47–60.

Norman, K. A., O'Reilly, R. C., & Huber, D. E. (2000). Modeling neocortical contributions to recognition memory. *The Cognitive Neuroscience Meeting, 2000*.

O'Keefe, J., & Nadel, L. (1978). *The hippocampus as a cognitive map*. Oxford, England: Oxford University Press.

O'Reilly, R. C. (1996). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation*, *8*(5), 895–938.

O'Reilly, R. C. (1998). Six principles for biologically-based computational models of cortical cognition. *Trends in Cognitive Sciences*, *2*(11), 455–462.

O'Reilly, R. C., & McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall: Avoiding a tradeoff. *Hippocampus*, *4*(6), 661–682.

O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. Cambridge, MA: MIT Press.

O'Reilly, R. C., Norman, K. A., & McClelland, J. L. (1998). A hippocampal model of recognition memory. In M. I. Jordan, M. J. Kearns, & S. A. Solla (Eds.), *Advances in neural information processing systems 10* (pp. 73–79). Cambridge, MA: MIT Press.

O'Reilly, R. C., & Rudy, J. W. (2000). Computational principles of learning in the neocortex and hippocampus. *Hippocampus*, *10*, 389–397.

O'Reilly, R. C., & Rudy, J. W. (in press). Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psychological Review*.

Phillips, R. G., & LeDoux, J. E. (1994). Lesions of the dorsal hippocampal formation interfere with background but not foreground contextual fear conditioning. *Learning and Memory*, *1*, 34–44.

Rolls, E. T. (1989). Functions of neuronal networks in the hippocampus and neocortex in memory. In J. H. Byrne, & W. O. Berry (Eds.), *Neural models of plasticity: Experimental and theoretical approaches* (pp. 240–265). San Diego, CA: Academic Press.

Rudy, J. W., & O'Reilly, R. C. (1999). Contextual fear conditioning, conjunctive representations, pattern completion, and the hippocampus. *Behavioral Neuroscience*, *113*, 867–880.

Rudy, J. W., & Sutherland, R. J. (1994). The memory coherence problem, configural associations, and the hippocampal system. In D. L. Schacter, & E. Tulving (Eds.), *Memory systems 1994* (pp. 119–146). Cambridge, MA: MIT Press.

Rudy, J. W., & Sutherland, R. W. (1995). Configural association theory and the hippocampal formation: An appraisal and reconfiguration. *Hippocampus*, *5*, 375–389.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*, 533–536.

Save, E., Poucet, B., Foreman, N., & Buhot, N. (1992). Object exploration and reactions to spatial and non-spatial changes in hooded rats following damage to parietal cortex or hippocampal formation. *Behavioral Neuroscience*, *106*, 447–456.

Schmajuk, N. A., & DiCarlo, J. J. (1992). Stimulus configuration, classical conditioning, and hippocampal function. *Psychological Review*, *99*(2), 268–305.

Sherry, D. F., & Schacter, D. L. (1987). The evolution of multiple memory systems. *Psychological Review*, *94*(4), 439–454.

Sloman, S. A., & Rumelhart, D. E. (1992). Reducing interference in distributed memories through episodic gating. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *Essays in honor of W. K. Estes* (pp. 227–248). Hillsdale, NJ: Erlbaum.

Sutherland, R. J., & Rudy, J. W. (1989). Configural association theory: The role of the hippocampal formation

in learning, memory, and amnesia. *Psychobiology*, *17*(2), 129–144.

Treves, A., & Rolls, E. T. (1994). A computational analysis of the role of the hippocampus in memory. *Hippocampus*, *4*, 374–392.

White, H. (1989). Learning in artificial neural networks: A statistical perspective. *Neural Computation*, *1*, 425–464.

Wickelgren, W. A. (1979). Chunking and consolidation: A theoretical synthesis of semantic networks, configuring in conditioning, S-R versus cognitive learning, normal forgetting, the amnesic syndrome, and the hippocampal arousal system. *Psychological Review*, *86*, 44–60.

Zipser, D., & Andersen, R. A. (1988). A backpropagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature*, *331*, 679–684.