6 Cortico-Hippocampal Interaction in Associative Learning

Chapter 4 argued that computational models of learning need to incorporate **stimulus representations** to allow appropriate generalization of learning between stimuli. The appropriate degree of generalization will depend on the particular problem, implying that representations should be adaptable to suit current task demands. However, the computational resources required to create appropriate new stimulus representations on the fly are considerable; neural-network researchers have addressed this problem by developing the error backpropagation algorithm described in chapter 4.

However, it is not clear that the sophisticated neural machinery needed to create the necessary new stimulus representations exists throughout the brain. One possible evolutionary alternative would be to localize some of the mechanisms for representational change in a central location (such as the hippocampus) so that other brain regions (such as cerebral cortex and cerebellum) could make use of these mechanisms as needed for particular tasks. This idea forms the basis for the two models of hippocampal function to be discussed in this chapter.

In both of these models, one network module representing the hippocampal region interacts with other network modules representing other brain regions, as in Marr's model (figure 5.9). *Hippocampal-region damage in these network models is simulated by disabling the hippocampal-region module and observing the behavior of the remaining modules.* These models can implement many aspects of associative learning, particularly classical conditioning, and they are useful for understanding how the hippocampal region may interact with the rest of the brain to facilitate certain kinds of learning.

The first model that we review, called the **cortico-hippocampal model,** is one that we ourselves originally developed to account for the effects of hippocampal-region damage on classical conditioning.[1] The basic idea of this model is that the hippocampal region is simulated as a predictive autoencoder that forms new internal-layer representations to compress redundant information while differentiating predictive information. These adaptive

representations are then adopted by long-term storage areas in the cortex and cerebellum.

The second model, called the **Schmajuk-DiCarlo model (**or **S-D model)** after its originators, takes a very different view of the hippocampal region. It presumes that the hippocampal region is necessary for the kinds of error-correction embodied in the Rescorla-Wagner model. Section 6.2 describes this model and its predictions, with a special emphasis on where it makes predictions that are similar to or divergent from those of our own cortico-hippocampal model.

In addition to these two computational models, there have been many qualitative, or noncomputational, theories of hippocampal-region function that seek to address many of the same behavioral phenomena. Section 6.3 reviews several prominent qualitative theories and shows how they relate to the computational models in sections 6.1 and 6.2.

Finally, section 6.4 describes one avenue of cognitive neuroscience research that has evolved out of the modeling work. As predicted by our cortico-hippocampal model, under some special conditions, individuals with medial temporal lobe (hippocampal-region) damage can actually learn simple associations faster than normal control subjects.

6.1 THE HIPPOCAMPAL REGION AND ADAPTIVE REPRESENTATIONS

Saul Steinberg created a famous cover for the *New Yorker* magazine, caricaturing his view of a typical New Yorker's mental map of the world. Manhattan was drawn in such fine detail that it took up most of the map. The rest of the country, the area between New Jersey and California, was squashed into a small area on the map, marked only by a farm silo and a few scattered rocks.

This painting satirizes many New Yorkers' belief that they are living in the most important place in the world. But it also illustrates an important psychological point. Fine distinctions that are meaningful to New Yorkers, such as the differences between Ninth and Tenth Avenues, are emphasized and highly differentiated in this mental map; these places are physically pulled apart and separated from surrounding areas. Broader distinctions that are irrelevant to the New Yorker, such as the difference between Kansas and Nebraska, are deemphasized or compressed and given less space in the map.

To some extent, we all create similar idiosyncratic worldviews with distorted representations; distinctions that are important to us are enhanced while less relevant ones are deemphasized. For example, students who are asked to sketch a map of the world tend to draw their home region disproportionately large and in the center of the map.[2] Figure 6.1 is such a map drawn by a student from Illinois, who overemphasized Illinois relative

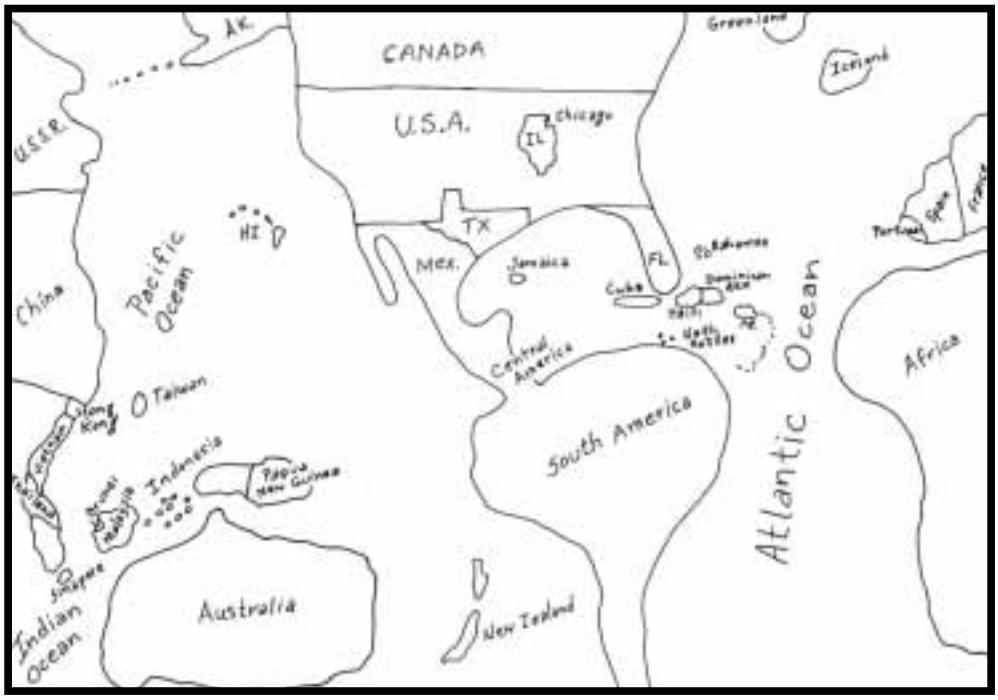


Figure 6.1 A student from Chicago, asked to sketch a map of the world, drew his home state disproportionately large while omitting most of the other states. North America was also disproportionately large in relation to the other continents. (Reproduced from Solso, 1991, Figure 10.11A, p. 289.)

to the rest of the country, omitted most states altogether, and enlarged North America relative to the other continents. Many American students show a similar pattern. In contrast, European students tend to draw Eurocentric maps, while students from Australia often place Australia and Asia in the center.

This kind of representational distortion, although somewhat comic in its egocentricity, is actually very useful. Memory is a limited resource, and individuals need to allocate that resource preferentially to items that are important to them. Thus, an experienced musician may actually devote more area of his brain to the fine control of finger movements than an average person would, while someone who has lost a hand through amputation will show shrinkage of the brain areas associated with finger movement. Chapter 8 will discuss these topics in more detail. For now, though, the chief question is: How can such representational changes come about? Who decides what kind of information is important enough to merit extra space and what is irrelevant enough to be shrunk like the Midwest on Steinberg's map?

Several years ago, we proposed a theory of hippocampal function in associative learning in which we argued that the hippocampal region is presumed to operate as an information gateway during the learning process.[3] Our theory assumes that the hippocampal region selects what information is allowed to enter storage and how it is to be encoded by other brain regions. Specifically, the theory argues that *the representation of redundant or unimportant information is shrunk, or* ***compressed,*** *while the representation of usefully predictive or otherwise meaningful information is elaborated, or* ***differentiated.*** According to this theory, the hippocampal region is critical for forming the kind of idiosyncratic maps of the world shown in figure 6.1. It turns out that this theory accounts for a range of behavioral data on representational processing with and without the hippocampus.

The Cortico-Hippocampal Model

Our theory assumes that the hippocampal region monitors statistical regularities in the environment and forms new stimulus representations that reflect these regularities. Specifically, *if two stimuli co-occur or make similar predictions about future reinforcement, their representations will be compressed to increase generalization between the stimuli. Conversely, if two stimuli never co-occur, and if they make different predictions about future reinforcement, their representations will be differentiated to decrease generalization between the stimuli.* The idea that the hippocampus can compress redundant information while differentiating predictive information is also consistent with the anatomy and physiology of the hippocampal region, as other researchers, such as William Levy, have noted previously.[4]

As described in chapter 5, a predictive autoencoder is capable of just this kind of function, compressing and differentiating representations in its internal layer. For this reason, our theory models the hippocampal-region network as a predictive autoencoder, mapping from stimulus inputs, through an internal layer, to outputs that reconstruct those inputs and also predict future reinforcement.

However, the hippocampal region is not the final site of memory storage, as is evidenced by myriad empirical data showing that old, well-established memories can survive hippocampal-region damage. Accordingly, *our model assumes that the representations developed in the hippocampal region are eventually adopted by other long-term storage sites in cortex and cerebellum*. Back in chapter 3, we noted that the cerebellar substrates of classical eyeblink conditioning are well characterized, and therefore we initially chose to apply our theory to this behavioral domain.

A modest elaboration of this cerebellar model introduced in chapter 3 is shown in figure 6.2A: Stimulus inputs (e.g., cues and context information) are processed by various primary cortical areas and then travel to the cerebellum via a structure called the pons. The cerebellum learns to map from these inputs to an output that drives a conditioned motor response. There is also an inhibitory feedback loop, through the inferior olive, that measures the error between the actual response (which is a prediction of unconditioned stimulus (US) arrival) and whether the US actually arrived. This allows the cerebellum to update connection strengths via an error-correcting procedure such as the Widrow-Hoff rule.

This simple cerebellar model does not make provision for any cortical learning, so we extended it into a hybrid cortico/cerebellar network as shown in figure 6.2B. Stimulus inputs (CSs and context information) are

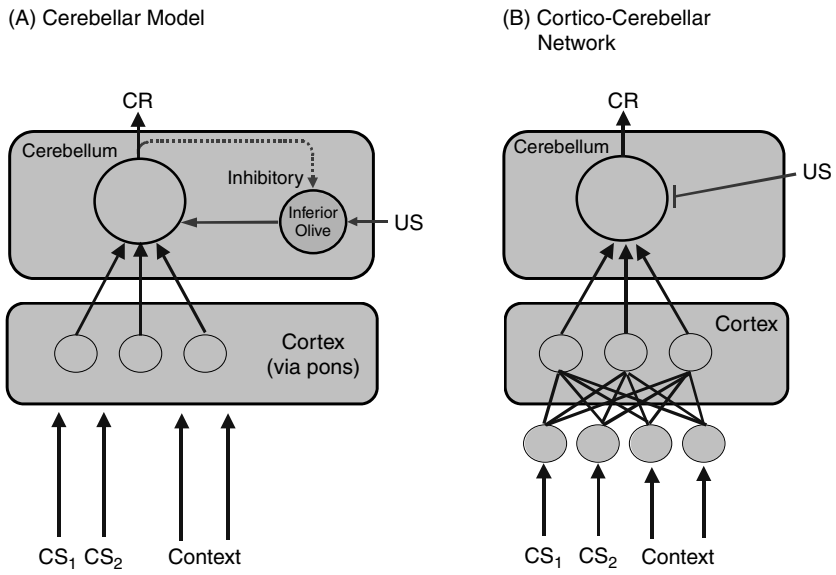


Figure 6.2 (A) The cerebellar model of chapter 3, redrawn to show preprocessing of sensory input by cortex; the cerebellum receives this input as well as direct CS information via a pathway through the pons. According to the model presented in chapter 3, the cerebellum learns to map from this input to a behavioral response (CR); the inferior olive computes the difference between US and CR and returns this error signal to guide learning in the cerebellum via the error-correcting Widrow-Hoff rule. (B) The same network, elaborated to allow an additional level of processing in the cortex. The upper layer of weights in this cortico/cerebellar network can still be trained by error correction to reduce the difference between CR and US; the lower layer of weights needs a training signal before it can use a second application of the error-correction rule.

provided as external input and activate an input node layer. Information travels through modifiable connections to an internal node layer representing cortical processing. In reality, of course, there might be an arbitrary number of successive cortical processing stages; for simplicity, we model only one.

The output of the cortical layer then travels to an output node representing the cerebellum, which integrates its sensory inputs and produces a conditioned response (CR). The difference between the CR and US is used as an error signal to drive learning in the upper layer of weights. The key question in multilayer network models—as discussed in chapter 4—is how one trains the lower layer of weights to alter internal-layer representations to facilitate learning.

This is where we believe that the hippocampal region plays a vital role. As shown in figure 6.3A, new representations formed in the hippocampal network's internal layer can serve as training signals for the cortico/cerebellar network's internal layer. In the simplest case, if each network has the same number of internal-layer nodes, the desired output for each internal-layer node in the cortico/cerebellar network is simply the actual output of the corresponding hippocampal-region network node.

Given this interpretation of the desired output for the cortico/cerebellar network's internal layer, the Widrow-Hoff error-correcting rule (from the Rescorla-Wagner model) can be used to train the lower layer of weights. Over many trials, the cortico/cerebellar network comes to adopt the same representations that were first developed in the hippocampal-region network.*

As figure 6.3B shows, hippocampal-region (HR) damage can be simulated by disabling the hippocampal-region network. In this case, *no new representations are acquired by the cortico/cerebellar network's internal layer. However, any previously acquired representations remain intact*. Thus, the cortico/cerebellar network can still learn new mappings from stimuli to responses based on its existing representations. For this reason, any new learning that does not require new representations (such as simply mapping one CS to a US) is likely to survive HR damage. However, any new learning that does depend

*If the total number of internal-layer nodes in the two networks is not equal, then the desired output for each cortico/cerebellar network internal-layer node may be some function of the activations of several nodes in the hippocampal-region network. The result is basically the same: The cortico/cerebellar network comes to adopt a linear transformation of the representations developed in the hippocampal-region network. This means that although there may be superficial differences in the two representations, they will show the same underlying logic: If the representations of two stimuli are compressed (or differentiated) in the hippocampal-region network, they will also be compressed (or differentiated) in the cortico/cerebellar network. See Gluck & Myers, 1993, for additional details.

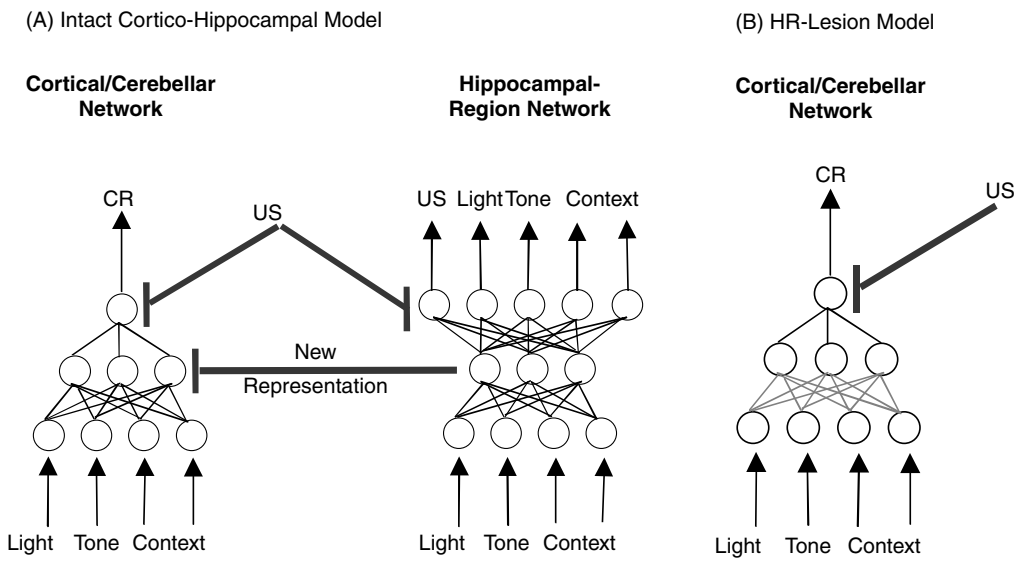


Figure 6.3 The cortico-hippocampal model. (A) The intact model receives inputs representing conditioned stimuli, such as lights and tones, as well as contextual information. One network, representing the processing that is dependent on the hippocampal region, learns to reconstruct these inputs and to predict the arrival of the unconditioned stimulus (US), such as a corneal air-puff. As it does, the hippocampal-region network forms new stimulus representations in its internal layer that compress redundant information and differentiate predictive information. A second network, assumed to represent long-term memory sites in cerebral and cerebellar cortices, adopts the internal representation provided by the hippocampal-region network and then maps from this to an output that represents the strength or probability of a conditioned response (CR), such as a protective eyeblink. (B) The HR-lesion model, in which the hippocampal-region network is disabled. The cortical network is no longer able to acquire new hippocampal-region-dependent internal representations, but it can still learn to map from existing representations to new behavioral responses.

on new representations (such as sensory preconditioning or configural learning) is expected by our model to be disrupted after HR damage. This distinction accounts for a great deal of data regarding HR-lesion effects, as we will describe below.

Representational Differentiation

As we noted earlier, the representations formed in the hippocampal-region network are subject to two biases: a bias to compress the representations of stimuli that are redundant and a bias to differentiate the representations of stimuli that predict different outcomes. Each of these biases can be used to

explain data in intact and HR-lesioned animals. First, we give several examples of learning behaviors that appear to involve representational differentiation.

Acquisition. The most rudimentary eyeblink conditioning task is **acquisition:** learning to respond to a cue that has been paired with the US. The Rescorla-Wagner model and the cerebellar model of figure 6.2 both capture this behavior, suggesting that the cerebellum alone should be sufficient to mediate conditioned acquisition—and hence learning should not be disrupted by HR lesion. Indeed, acquisition of a conditioned eyeblink response is not disrupted by HR lesion in humans (figure 6.4A), rabbits (figure 6.4B), or rats.[5]

Conditioned acquisition is simulated in the intact cortico-hippocampal model by presenting a series of training trials. First, the model is given trials consisting of just the experimental context—call it X—a series of inputs

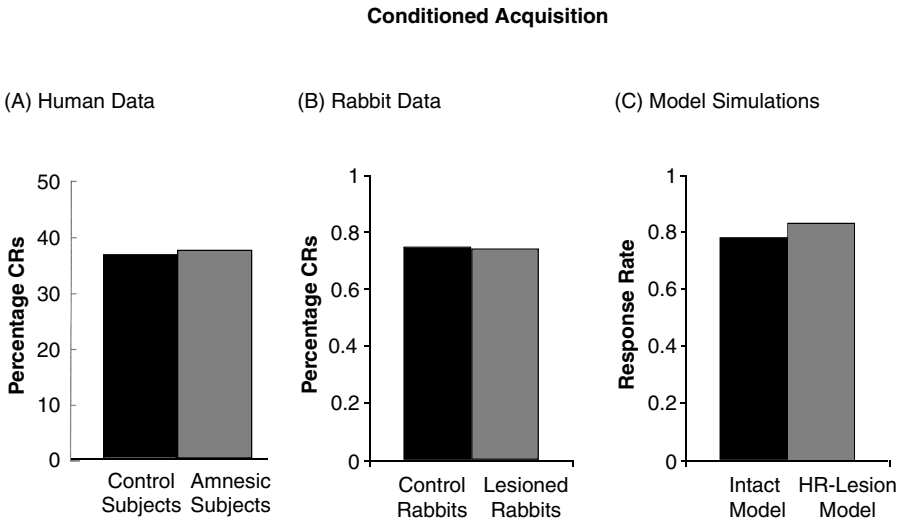


Figure 6.4 Conditioned acquisition, learning that a tone CS predicts an airpuff US, is not disrupted by hippocampal-region damage. (A) Humans with medial temporal lobe damage, including hippocampal-region damage, show normal eyeblink conditioning (Gabrieli et al., 1995). (B) Rabbits with hippocampal-region damage similarly acquire the conditioned eyeblink response as fast as control rabbits (Solomon & Moore, 1975). (C) Similarly, the intact cortico-hippocampal model and HR-lesion model learn at the same speed (Myers et al., 1996). For all graphs, response rate and percentage CRs represent proportion of trials generating CRs after a fixed number of CS-US pairings. (Adapted from Myers, Ermita, et al., 1996, Figure 4.)

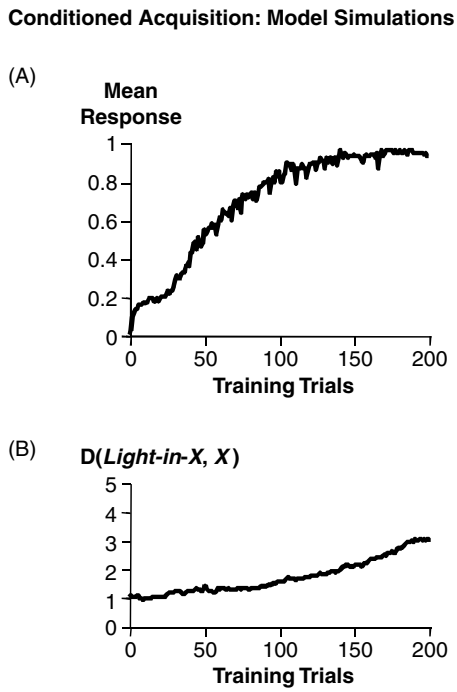


Figure 6.5 Conditioned acquisition in the intact cortico-hippocampal model. (A) Learning curve. (B) Representational differentiation of the light CS from the context X alone during training, reflected in increasing *D*(light-in-X, X).

meant to represent the background sights, smells, and sounds of the experimental setup; the model learns not to give a conditioned response to the context alone. These trials correspond to the time spent acclimating an animal to the experimental chamber, before any explicit training begins, a standard procedure in experimental studies of animal conditioning.

Next comes the actual acquisition training. Because the training takes place in context X, learning to respond to a light CS can be redefined as learning to respond to light-in-X but not to the context alone X−. With enough training, the model learns to respond when the light is present but not to the context alone, as is seen in figure 6.5A.* Figure 6.5B shows the corresponding changes in internal-layer representation in the cortico/cerebellar network, copied from the representations in the hippocampal-region network. The

*All figures that show performance of the cortico-hippocampal model are the average of ten simulation runs. Error bars on simulation data reflect variance among multiple simulation runs. Full details of the model presented in section 6.1 are given in Appendix 6.1 at the end of this chapter.

difference in representation between light-in-X and X alone—*D*(light-in-X, X), as defined in the previous chapter—grows gradually but consistently with increasing training. This representational differentiation facilitates the cortico/cerebellar network's task of mapping the two inputs to different responses.

However, this new differentiated representation is probably not necessary to acquire a conditioned response to a single light CS. The task is so simple that just about any random recoding in the lower layer of cortico/cerebellar network weights is probably sufficient. As long as there is at least one node in the internal layer that gives a different response to light-in-X and X alone, that node can be used to drive the presence or absence of a CR. In fact, the HR-lesioned model can learn the correct response about as quickly as the intact model (figure 6.4C). Thus, the cortico-hippocampal model correctly accounts for the finding that HR lesion does not impair acquisition of a simple CS-US association.

Discrimination and Reversal. Simple **discrimination** involves learning that one CS (light+) predicts the US while a second CS (tone−) does not. This means that conditioned responses should follow light+ but not tone−. In general, discrimination learning in the eyeblink-conditioning paradigm is not disrupted by hippocampal-region damage (figure 6.6A).[6] Similarly, hippocampal-region damage generally does not impair a range of discrimination tasks in animals, including discrimination of odors, objects, textures, and sounds.[7]

In the intact cortico-hippocampal model, the hippocampal-region network constructs new representations that differentiate light+ and tone−, facilitating the mapping of light+ to one response and tone− to another (figure 6.6C,D). However, the discrimination task is so simple that such representational changes are probably not necessary; any random initial representations in the cortico/cerebellar network are probably different enough to allow mapping to different responses. Thus, the HR-lesioned model should be able to learn a conditioned discrimination. As is shown in figure 6.6B, there is indeed no impairment: The HR-lesioned model reaches criterion performance just as quickly as the intact model.

These empirical data have often been interpreted as arguing that conditioned discrimination is hippocampal-independent. Our model offers a different interpretation: *The hippocampal-region may not be strictly necessary for some simple kinds of learning; but when it is present, it normally contributes to all learning*. Even in a simple task such as discrimination (or acquisition), where *a priori* representations probably suffice to allow learning, the hippocampal

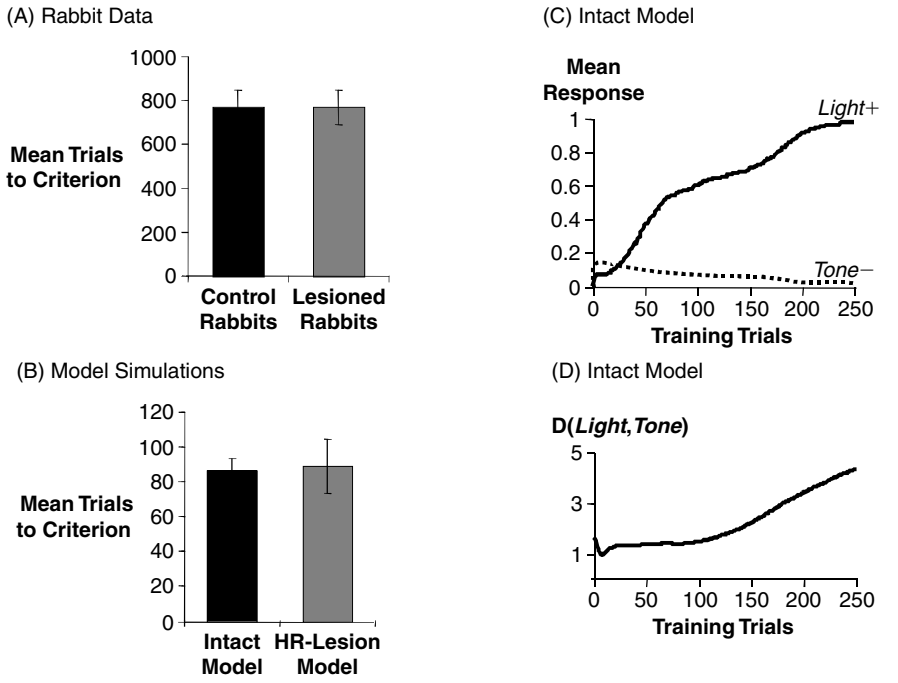


Figure 6.6 Conditioned discrimination: learning to respond to one CS light+, which is paired with the US, but not to CS tone−, which is not paired with the US. (A) Control rabbits learn this task in about 800 trials; rabbits with hippocampal-region lesion learn at the same speed (Berger & Orr, 1983). (B) The intact and HR-lesion cortico-hippocampal model likewise learn at the same speed. (C) In the intact model, there is an initial period when the model responds weakly but equally to both light+ and tone−; then the model begins to discriminate. (D) During learning, the representations of light and tone are differentiated, reflected in increasing *D*(light, tone).

region is constantly forming new stimulus representations that compress redundant information while differentiating predictive information, whether these new representations are needed or not.

However, the usefulness of this hippocampal participation becomes apparent if task demands change. For example, suppose the discrimination is reversed so that after learning to respond to light+ but not tone−, the contingencies are reversed, so tone+ now begins to predict the US and light− does not. In our intact model, the hippocampal-region network has already done the work of differentiating the representations of light and tone; once the contingencies reverse, all that needs to be done is to map those representations to new responses. In the lesioned model, the situation is quite different: The representations of light and tone are fixed, and so they are not differentiated during the original discrimination. Thus, the reversal

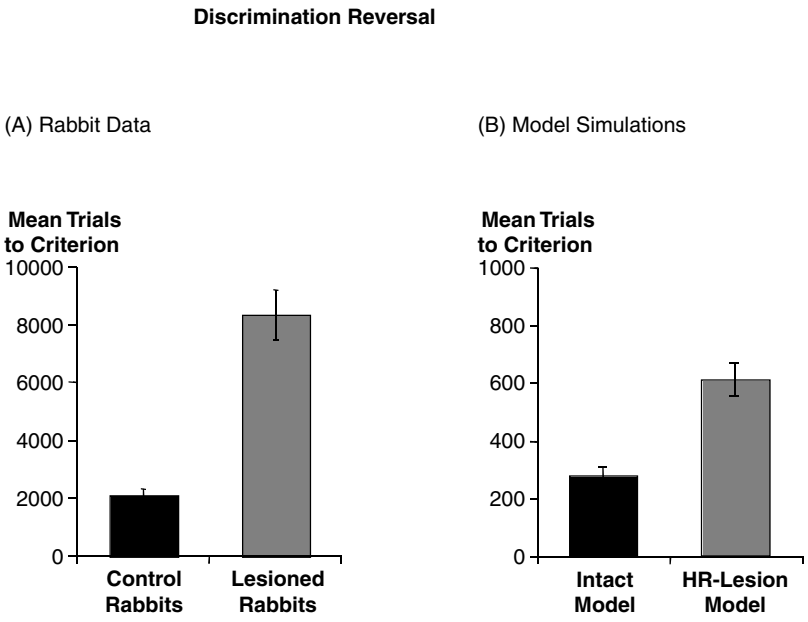


Figure 6.7 Discrimination reversal. (A) In the rabbit eyeblink preparation, reversal of a conditioned discrimination is strongly impaired in animals with hippocampal lesion. (Plotted from data presented in Berger & Orr, 1983.) (B) Similarly, the HR-lesion model is slower to reverse than the intact cortico-hippocampal model.

requires first unlearning the original discrimination and then learning the reversed discrimination. This process may be quite lengthy in comparison to reversal in the intact model (figure 6.7B). In rabbit eyeblink conditioning, several studies show that hippocampal-region damage disrupts discrimination reversal (figure 6.7A).[8]

Other Behaviors Involving Predictive Differentiation. Our cortico-hippocampal model also predicts that many other paradigms that involve representational differentiation will be disrupted after HR lesion.[9] These tasks include **easy-hard transfer** (the finding that learning a hard discrimination is facilitated by prior training on an easier version of the task), the **overtraining reversal effect** (the finding that reversal is speeded if the original discrimination is trained for many days beyond criterion performance), and **nonmonotonic development of the stimulus generalization gradient** (the finding that learning about one stimulus early in training generalizes strongly to other stimuli, while generalization is reduced when similar learning takes place later in training).

These are novel predictions: no data currently exist documenting the behavior of HR-lesioned animals on these tasks. However, experiments in our lab at Rutgers-Newark are under way to test some of these predictions. The results of such experiments will provide a critical test of the model, either confirming its predictions or showing that the model requires revision.

Representational Compression

Just as the hippocampal region is assumed to differentiate the representations of stimuli that should be mapped to different responses, the hippocampal region is assumed to compress the representations of stimuli that co-occur and should be mapped to similar responses. Behaviors that reflect representational compression should be disrupted after hippocampal-region damage.

Sensory Preconditioning. One simple example is **sensory preconditioning,** which was discussed in several earlier chapters. Recall that sensory preconditioning involves unreinforced exposure to a compound of two stimuli (tone&light− exposure), followed by light-US pairings (light+ training). The associations learned to the light should partially transfer to tone, as a result of the paired exposure. Hippocampal-region damage (specifically fimbrial lesion) abolishes sensory preconditioning in the rabbit eyeblink preparation, as shown in figure 6.8A.*[10] Because the predictive autoencoder in chapter 5 was sufficient to mediate this effect (refer to figure 5.19), it should come as no surprise that our intact cortico-hippocampal model shows the same effect (figure 6.8B).[11] In the intact model, tone&light− exposure results in compression of the representations of tone and light, since both stimuli co-occur and neither predicts the US or any other salient event. Subsequent associations to light partially activate the representation of tone, and the learning transfers. In the lesioned model, there are no representational changes during the exposure phase, and as long as light and tone are distinct stimuli that activate different (fixed) representations, there is little chance that associations made to light will transfer to tone.

*Fimbrial lesion does not involve removal of the hippocampus but instead cuts an important input and output pathway by which subcortical structures communicate with the hippocampus. As such, the effects of fimbrial lesion may not be identical to hippocampal lesion. The implications of damage or disruption to this pathway will be further discussed in chapter 9.

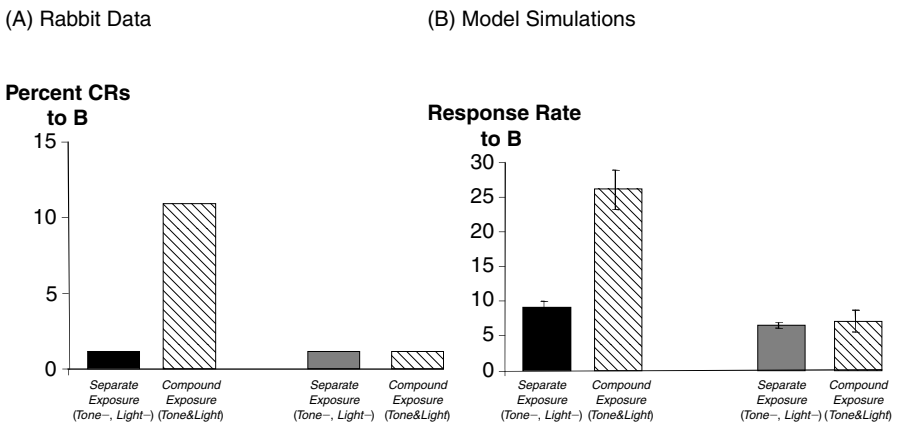


Figure 6.8 Sensory preconditioning. (A) In rabbit eyeblink conditioning, phase 1 exposure to the compound tone&light−, followed by light+ training, results in stronger phase 3 responding to tone than in animals that are given separate exposure to the components (tone−, light−) in phase 1; this effect is eliminated in animals with damage to the hippocampal region (fornix lesion). (Adapted from data presented in Port and Patterson, 1984.) (B) Similarly, the intact but not HR-lesioned cortico-hippocampal model shows sensory preconditioning.

Learned Irrelevance. Another behavior involving representational compression is **learned irrelevance.**[12] The paradigm is schematized in table 6.1. In phase 1, subjects in the exposed group are given presentations of a CS (e.g., light) and a US, uncorrelated with each other. Subjects in the nonexposed group are given equivalent time in the experimental context but receive no presentations of light or the US. In phase 2, all subjects receive light-US pairings. As shown in figure 6.9A, subjects in the exposed group are much slower to learn the light-US association.

In the intact cortico-hippocampal model, phase 1 exposure to a CS (e.g., light) and a US causes representational changes. The representation of the light becomes compressed, together with the representations of the background contextual cues, since neither predicts the US well. In effect, the light is treated as a sometimes-occurring aspect of the context, one that is of no use in predicting US arrival. This representational compression of light and context will hinder phase 2 learning to respond to the light but not the context alone. Thus, as shown in figure 6.9B, there is a learned irrelevance effect in the intact cortico-hippocampal model. Since learned irrelevance is interpreted in terms of representational compression, it is not shown in the HR-lesion model (figure 6.10B).

Table 6.1 The Learned Irrelevance Paradigm

Group	Phase 1	Phase 2
CS/US exposure	Light and airpuff (uncorrelated)	Light → airpuff . . . . *SLOW!*
Sit exposure	(Animal sits in experimental chamber)	Light → airpuff . . . . *normal speed*

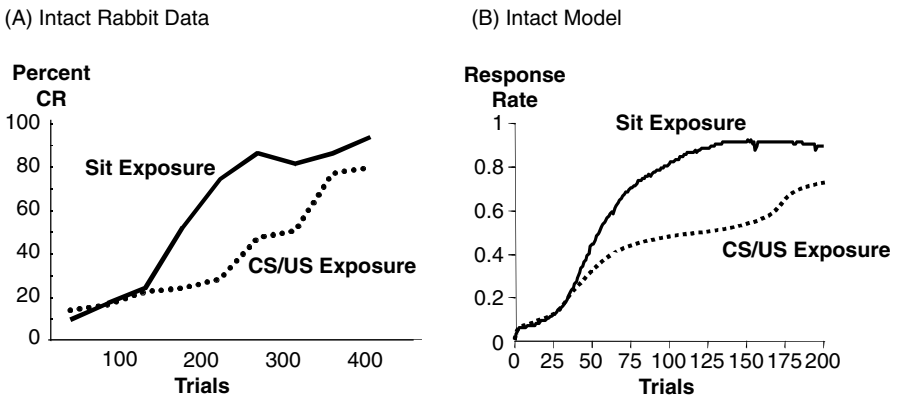


Figure 6.9 Learned irrelevance. Both intact rabbits (A) and the intact cortico-hippocampal model (B) show slower learning of a CS-US association following exposure to the CS and US uncorrelated with each other. (A is drawn from data presented in Allen, Chelius, & Gluck, 1998.)

At the time we first published this prediction, it was a novel implication of the cortico-hippocampal model. Since then, we have shown in our lab that learned irrelevance is severely disrupted by HR lesion in the rabbit eyeblink-conditioning preparation (figure 6.10A).[13] These results provide further evidence that the representational compression contributing to learned irrelevance depends on the hippocampal region. Our empirical data further demonstrate that the exact lesion extent is critical in determining whether or not learned irrelevance is disrupted, suggesting that different hippocampal-region structures contribute differently to the effect. We will return to this issue, and the actual empirical data, in chapter 9.

Our cortico-hippocampal model predicts that other behaviors that reflect representational compression will also be disrupted by HR lesion. These include **latent inhibition** and **contextual effects,** which will be discussed more fully in chapter 7. Some of these predictions have also been confirmed by recent experimental studies.

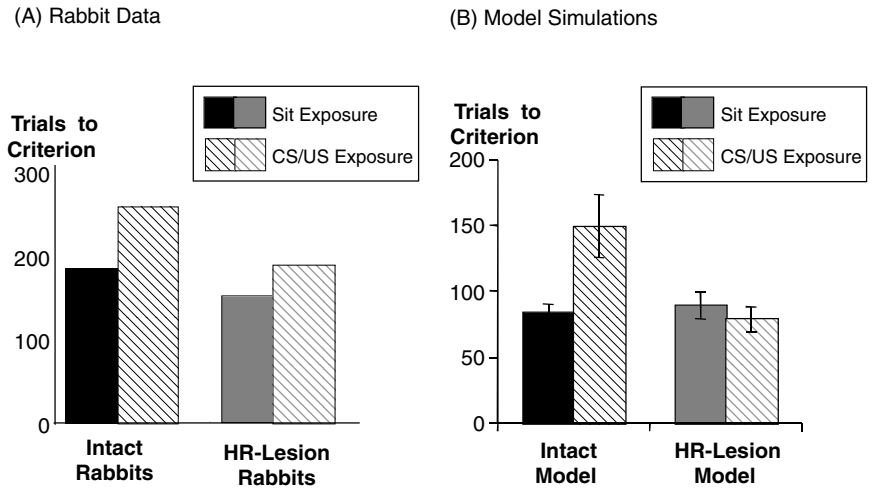


Figure 6.10 Learned irrelevance and HR-lesion. (A) In rabbits, hippocampal-region damage (specifically, entorhinal lesion that also cuts off the major information pathways into and out of hippocampus) eliminates learned irrelevance: Animals that are given prior exposure to the CS and US uncorrelated learn at about the same rate as animals that are given exposure to the context only (Sit Exposure). (B) Likewise, the HR-lesion model does not show learned irrelevance. (A is drawn from data presented in Allen, Chelius, & Gluck, 1998.)

Limitations of the Cortico-Hippocampal Model

Although our cortico-hippocampal model accounts for a considerable range of data on intact and hippocampal-lesioned animals, there are several effects that it does not address. The model is specifically limited to model classical conditioning. This limited domain was chosen specifically because it is possible to construct a model of the cerebellar substrates of eyeblink conditioning with some assurance that the model accurately reflects the brain substrates. However, this approach means that the model does not apply easily to other domains. Recently, we have shown that the model can be extended to address **instrumental conditioning,** a form of learning in which reinforcement is contingent on the subject's response, and **category learning,** in which people learn to classify objects into predefined classes.[14] However, there are other domains that lie well outside the model's ability; these include spatial learning, declarative memory, and delayed nonmatch to sample (DNMS), three behaviors that were mentioned in chapter 2 as important areas of research into hippocampal function. These are limitations of the current computational model but not necessarily of the basic underlying theory. An important direction for future research will be to see how well the

fundamental information-processing principles of representational compression and differentiation can account for hippocampal function in more complex behaviors such as spatial learning and recognition learning.

Extinction. There are, however, phenomena within the model's domain of classical conditioning for which the model fails to accurately account for all relevant empirical data. One example is **extinction.** After learning a response to one CS (light+ training), if an animal is given CS-alone trials (light− training), the conditioned response gradually weakens, or extinguishes. Extinction is problematic as a psychological phenomenon because there is considerable evidence that the process is more complicated than simply undoing a CS-US association; instead, subjects appear to learn a CS-noUS association that competes with the earlier CS-US association.[15] The idea that the earlier CS-US association is not destroyed, but merely suppressed, is consistent with the finding that an extinguished association can be reacquired more quickly than it was originally acquired.[16]

The cortico-hippocampal model does not contain any explicit mechanism for the simultaneous maintenance of CS-US and CS-noUS associations. During extinction, all that occurs is that the mapping between the representation of CS and the US is replaced by a mapping to noUS. No representational changes are involved, and so extinction occurs at the same speed in the intact and lesioned models (figure 6.11B). This behavior is superficially correct—hippocampal-region damage does not affect extinction in animals (figure 6.11A)[17]—but is nonetheless an oversimplification. Extinction also occurs much more quickly in the model than in experimental subjects.

To account for these data, additional mechanisms would have to be postulated in the model to account for all these aspects of extinction. At the present time, sufficient controversy surrounds the true nature of extinction that it seems premature to try to add such a mechanism to the model. However, it would be a fruitful exercise to try to implement some of the possible mechanisms in a model of extinction and see which accounts for the greatest array of empirical data. It has been suggested that extinction does not erase CS-US learning but, rather, makes this learning more sensitive to context.[18] Thus, the correct response in the old context was to produce a CR, but in the new context, no CR should occur. We will return to contextual issues in chapter 7.

Timing Effects. A second major class of data for which our original cortico-hippocampal model does not account is stimulus interval effects such as **trace conditioning.** Trace conditioning is defined by introducing a short interval between CS offset and US onset (refer to figure 2.17). Trace

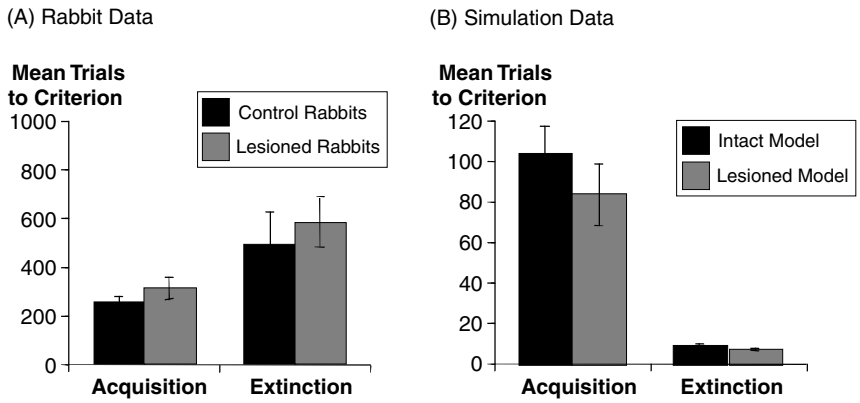


Figure 6.11 Acquisition and extinction. (A) In the rabbit eyeblink preparation, extinction of a trained response is not significantly slowed by hippocampal-region damage. In both cases, extinction typically takes longer than acquisition of the original response. (Plotted from data presented in Berger & Orr, 1983.) (B) In the cortico-hippocampal model, extinction is not slowed in the HR-lesion model; however, in contrast to the animal data, both intact and HR-lesioned models show extremely rapid extinction. The model probably does not adequately capture the complexities of extinction in animals.

conditioning is disrupted by hippocampal-region damage—if the trace interval is long enough.[19] The version of the cortico-hippocampal model described here does not incorporate temporal information such as the number of milliseconds between CS and US onset. Therefore, it cannot directly model stimulus interval effects. Recently, we have developed a generalized version of our model that includes recurrent connections within the network, thereby allowing it to demonstrate some aspects of temporal and sequential processing, including trace conditioning.[20]

However, introducing temporal information into the cortico-hippocampal model does not solve a more fundamental problem. The cortico-hippocampal model assumes that the hippocampal region is critical when new stimulus representations are required that involve redundancy compression or predictive differentiation. There is no obvious way to relate this assumption to the finding of a hippocampal role in trace conditioning; trace conditioning does not seem to require new stimulus representations—only the formation of an association between CS and US. Thus, our theory does not provide a good understanding of why trace conditioning should depend on the hippocampal region. At present, the best that can be done is to note that the model does not rule out additional possible hippocampal-region functions such as a role in short-term memory that might help to maintain a representation of the CS during a trace interval. Interestingly, trace

conditioning is not universally reported to depend on the hippocampal region.[21] One possible reason for this discrepancy between studies is variance in exact lesion extent;[22] this would be consistent with a suggestion that different hippocampal-region subareas perform different functions and that some are more involved in trace conditioning than others.

Many other computational models do address a role for the hippocampus in timing (including trace conditioning).[23] Future work remains to determine whether a single model can capture both the representational and temporal aspects of hippocampal-region function (and also spatial learning, DNMS, and episodic memory).

Neurophysiological Support for the Cortico-Hippocampal Model

If the hippocampal region does adapt stimulus representations to minimize redundancy while preserving predictive information, it should be possible to observe the results of this process via neurophysiological recordings of cell firing in the hippocampal region. It is now possible to simultaneously record the activation of dozens of neurons within a small area of brain. The set of firing activity across a set of neurons is a pattern analogous to the activities across a set of nodes in a network model and can be viewed as the brain's representation of the current inputs. The difference between the brain's representation of one input A as a pattern of neuronal firing and a second input B can be quantified by using a D(A, B) metric just like that defined for the network model. Thus, it is possible to measure D(A, B) in a specific region of the brain both before and after training and then calculate whether the neural representations have become more or less similar. This approach has obvious limitations: It is possible to simultaneously record only a small sample of the vast number of neurons in any brain region, and—as with any sampling method—it is possible that the small sample may not accurately reflect the larger population. However, if the predicted changes in D(A, B) are visible among even a small sample of neurons, it is a reasonable inference that similar changes occur throughout the population.

By using this method, it is possible to observe changes in hippocampal firing patterns as a result of learning. For example, during rabbit eyeblink conditioning, some neurons in the hippocampus that do not respond strongly to a CS will gradually increase responding to that CS if it is repeatedly paired with a US.[24] In a discrimination task, some neurons will respond to the CS (light+) but not to the CS (tone−).[25] Averaged over many neurons, these changes will result in increased D(light, tone) much like that shown in figure 6.6D. Importantly, these changes in hippocampal neuronal activity *precede*

the development of the behavioral CR,[26] just as representational changes precede development of the response in the predictive autoencoder of our hippocampal model (refer to figure 5.18).

Rabbits are not the only animals to show these kinds of changes predicted by our cortico–hippocampal model. In one experiment, monkeys were trained on a visual discrimination, in which they had to make an arm movement when they saw one stimulus A+ but not when they saw a second stimulus B−. Figure 6.12A shows how a single neuron in the hippocampus responded during this task.[27] Initially (trials 1 through 20), there is a similar response to both stimuli A and B. With further training, these neurons begin to become more active when stimulus A was presented than when stimulus B was presented. These changes preceded the behavioral evidence of learning, evidenced by correct arm movements. Of all the hippocampal neurons

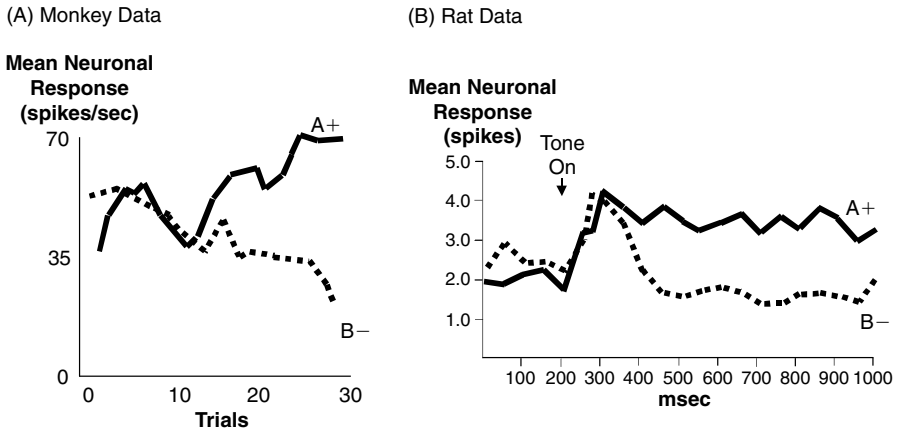


Figure 6.12 Representational changes during discrimination learning. Neurophysiological recordings of neuronal activity suggest representational differentiation during learning. (A) Recordings from monkey hippocampus during learning to make a motor response to one stimulus (A+) but not a second (B−). Initially (trials 1–10), there is little difference in responding to the two stimuli; with further training (trials 20–30), there is significantly greater neuronal activity to the rewarded stimulus A+. (Adapted from Cahusec et al., 1993, Figure 2.) (B) Recordings from rat dentate gyrus during training to respond to tone A+ but not tone B−. The graph shows the pattern of responding on a single presentation of the tones, averaged across several trials, in an animal that had learned the appropriate responses. When either tone comes on (200 msec into the trial), there is initially a response. However, only for the rewarded stimulus A+ is this response maintained for the duration of the trial. (Adapted from Deadwyler, West, & Lynch, 1979, Figure 4.)

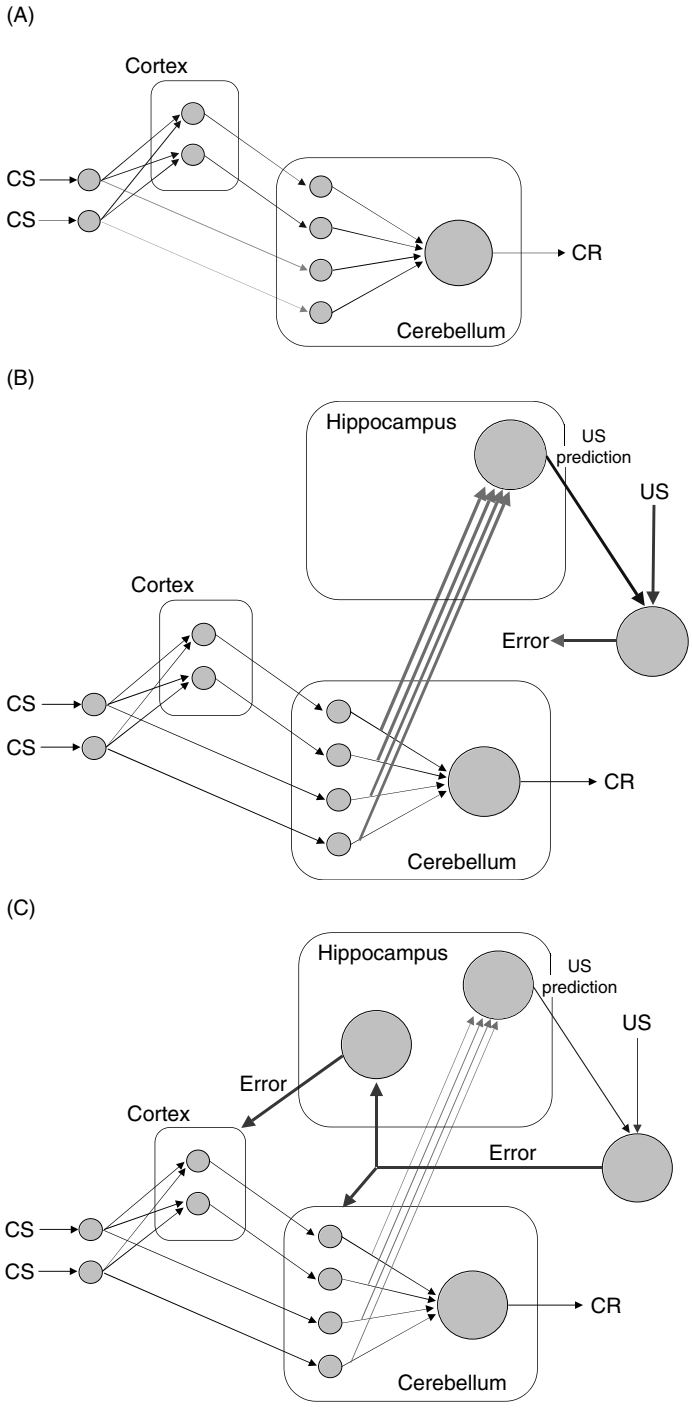
that showed task-related activity, about 22% altered their activity patterns to discriminate the two stimuli; most of these maintained these new response patterns across the experiment. However, many did not maintain their new activity patterns once the behavior was fully acquired. As the authors of this study note, this would be consistent with the idea that the hippocampus has a limited capacity for pattern storage, and at some point, old learning is transferred elsewhere (e.g., cortex) and new learning overwrites the old in hippocampus.[28]

In another study, Deadwyler and colleagues recorded neuronal activity in the dentate gyrus of rats that were being trained to make a motor response to one tone A and not to another tone B.[29] Initially, neuronal activity (in the form of extracellular unit discharge patterns) looked similar after presentation of either tone. By the time the conditioned response had been acquired, neuronal discharge in the dentate gyrus differentiated the two stimuli; specifically, neurons might respond to both stimuli, but only the rewarded stimulus (A) elicited sustained activity (figure 6.12B). Although these tone-evoked responses occur slightly before the initiation of the behavioral response, Deadwyler et al. argue that the dentate activity is probably not directly related to the production of a motor movement. Instead, the dentate gyrus may contribute to learning which of two or more competing responses is appropriate to a given stimulus and may encode information about expected reward by means of differential discharge patterns. A related study has also shown that hippocampal cells in the rat that encode place information change as a result of learning, differentiating the representations of landmarks.[30]

Taken together, *the neurophysiological evidence currently available is remarkably consistent with the implications of our cortico-hippocampal model, suggesting that hippocampal neuronal representations can and do change to reflect associations between stimuli and rewards*. The results do not prove that the hippocampus itself creates these representational changes; it is possible that another brain region develops appropriate representations and merely passes this information to hippocampus. However, these neurophysiological findings are clearly consistent with the idea that the hippocampus creates new stimulus representations, much like the internal-layer nodes in a predictive autoencoder.

6.2 SCHMAJUK AND DICARLO (S-D) MODEL

Nestor Schmajuk and his colleagues have presented in several papers an evolving series of computational models of cortico-hippocampal interaction in conditioned learning.[31] These models are similar in spirit and aim to our cortico-hippocampal model in that these models are concerned with



information-processing roles for different brain subregions and how these subregions exchange information. Moreover, Schmajuk's models are also meant to address much of the same body of empirical data as our cortico-hippocampal model. However, the particular function that Schmajuk and his collaborators assign to the hippocampus differs from that supposed by our model, leading to some predictions that may differentiate the two models.

Schmajuk and DiCarlo have presented a computational model of the hippocampal-system processing in classical eyeblink conditioning, often referred to as the **S-D model.**[32] *The S-D model assumes that CS information reaches the cerebellum via two routes: a direct path and an indirect path involving association cortex (figure 6.13A)*. This dual-pathway assumption seems anatomically valid.[33] Schmajuk and DiCarlo suggest that CS information is combined in the cortex to allow configural learning. Then, they argue, the cerebellum integrates information from both the direct and indirect pathways to produce a CR.

The S-D model assumes that the hippocampal system has two roles in conditioned learning. The first is shown in 6.13B. According to the Rescorla-Wagner learning rule (refer to MathBox 3.1), learning is proportional to an error measure, which is the difference between the actual and predicted US. The S-D model assumes that the hippocampus is critical in computing this error. Specifically, *the hippocampus in their model is presumed to calculate the US prediction; other brain areas compare this predicted US against the actual US and calculate the total error*. This error signal is then used to guide learning.

Figure 6.13 (A) The S-D model of classical eyeblink conditioning (Schmajuk & DiCarlo, 1992) assumes that CS information reaches the cerebellum by two pathways: a direct CS-cerebellum pathway and an indirect pathway via neocortex. This assumption is consistent with known anatomy. The S-D model assumes that the neocortex recombines CS information to allow configural learning. Finally, the cerebellum integrates information from both the direct and indirect CS pathways to produce a CR. (B) The S-D model assumes that the hippocampus has two roles in conditioned learning. First, it calculates the strength of the US prediction by summing the activations from cerebellum. The hippocampal output is a measure of how strongly the US is predicted. This US prediction measure is passed to other brain regions, which compare it against the actual US and compute a prediction error. This prediction error is similar to the error measure in the Rescorla-Wagner rule, and the learning rate is proportional to the magnitude of this error. (C) The second function of the hippocampus in the S-D model is to broadcast the error signal to the neocortex. According to this model, hippocampal-region damage should impair both the computation of the aggregate error signal as well as the ability of the neocortex to develop new configural nodes.

Figure 6.13C illustrates the second role for the hippocampal region in the S-D model: *The cerebellar units can update weights directly on the basis of the error signal, whereas the cortical units require specialized error signals broadcast by the hippocampus*. In neural network terms, the cortical units are hidden units between the input and output layers, and the hippocampal circuitry is specialized to compute error signals for these hidden units—by implementing a version of error backpropagation.

Hippocampal-system damage is simulated in the S-D model by disabling *both* of the putative hippocampal functions. First, there is no longer any way to calculate the predicted US. Most of the power of the Rescorla-Wagner model (and the Widrow-Hoff learning rule) comes from the ability to predict the US on the basis of *all* available cues. The lesioned S-D model cannot form such an **aggregate prediction** of the US and is reduced to simple learning about individual CSs. Thus, the S-D model predicts that hippocampal damage will impair behavioral phenomena such as blocking that require this error signal based on an aggregate prediction of the US.

Second, the lesioned S-D model cannot form new configural nodes in the cortex, and it is restricted to simpler CS-US learning in the cerebellum. Thus, the S-D model correctly produces unimpaired CS-US learning after hippocampal lesion: The indirect CS-cortex-cerebellum pathway is dysfunctional because the cortical units cannot update without hippocampal error signals, but the direct CS-cerebellum pathway is operational and allows learning. Other forms of learning, such as sensory preconditioning, that depend on CS-CS associations, are disrupted by damage either to the cortical system or to the hippocampal system that provides its error signals.

There are numerous additional complexities to the S-D model that are not discussed here,[34] but the description given above is sufficient to illustrate the basic principles by which the model operates. Applied to classical conditioning, the S-D model can account for a sizable range of empirical findings. These include the sparing of discrimination learning, but impairment of reversal, after hippocampal lesion; the broadening of the generalization gradient in lesioned animals; and the loss of latent inhibition after hippocampal lesion.[35] In addition, because it is a real-time model, it can successfully account for such temporal effects as trace conditioning and phasic cue occasion setting, which are beyond the scope of our own cortico-hippocampal model.[36]

In later work, the S-D model has been extended to include attentional processing and novelty detection.[37] The basic idea behind these subsequent models is that when a mismatch occurs between prediction and reality, attention to the current stimuli is increased and the prediction generator is updated. Thus, during the first CS-US pairing, the unfamiliar CS generates

high attention, facilitating its association with the US; during latent inhibition, the CS becomes familiar, reducing attention and impairing the ability of that CS to enter into subsequent associations. More recently, Schmajuk has suggested that the two hippocampal-system functions proposed in the S-D model can be subdivided and mapped to various hippocampal-region substructures.[38] Chapter 9 will discuss these elaborations in more detail.

Comparison with Gluck and Myers's Cortico-Hippocampal Model

Because the S-D model and our cortico-hippocampal model both address the same domain (classical eyeblink conditioning), there is a great deal of overlap in their predictions. However, there are several important points on which the two models differ. The most basic of these concerns the ability to predict the US based on all available CSs. The S-D model sites this aggregate prediction in the hippocampus and assumes that hippocampal lesion abolishes it. Our cortico-hippocampal model, by contrast, follows Thompson's cerebellar theory[39] in assuming that this (and all other aspects of the Rescorla-Wagner model) are implemented in the cerebellum. The HR-lesioned cortico-hippocampal model does continue to compute a prediction of the US based on all available cues and to use this information to guide error-correction learning. Thus, the S-D model and cortico-hippocampal model make different predictions about the effects of hippocampal-region damage on behavioral effects that, in the intact animal, reflect aggregate prediction of the US. While there are behavioral data testing these predictions, the data are unfortunately mixed in many cases.

Conditioned Inhibition. One example of such an effect is **conditioned inhibition:** learning to respond to one cue (e.g., light+) when presented alone but not when paired with another cue (e.g., tone&light−).

The Rescorla-Wagner model can solve this task by setting a positive weight on light and a negative weight on tone so that light alone produces a response but tone and light together cancel each other out and produce no response. Both the Rescorla-Wagner model and the cerebellar model of chapter 3 show this effect. This implies that the cerebellum should be sufficient to mediate conditioned inhibition, and therefore hippocampal-region damage should not disrupt performance. Paul Solomon tested this idea in rabbit eyeblink conditioning and found that rabbits with hippocampal-region damage could learn the task just as well as control rabbits, as shown in figure 6.14A.[40]

Because the cortico/cerebellar network of figure 6.2B incorporates the error-correcting learning procedure from the Rescorla-Wagner model, both

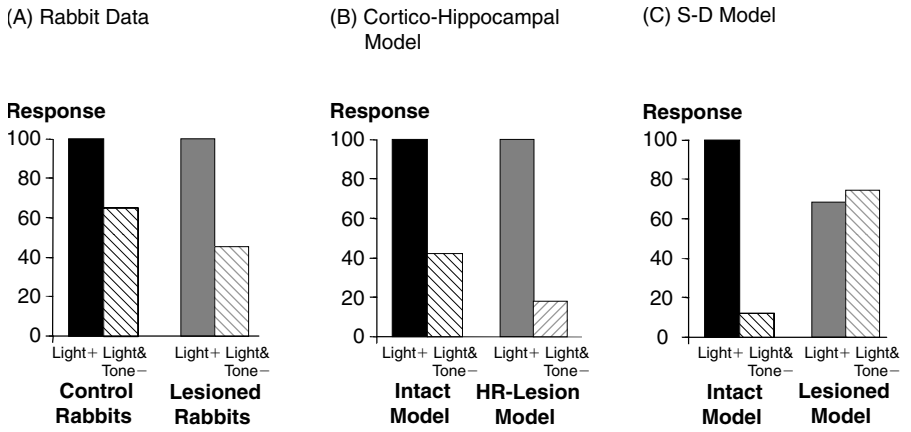


Figure 6.14 Conditioned inhibition: learning to respond to light+ but not light&tone−. Results are expressed as a percent of response to light+ in intact animals or model. (A) Rabbit eyeblink conditioning data: Both intact and hippocampal-lesioned animals can learn a strong response to light+ and a weak response to light&tone− after about 700 blocks of training; the lesioned animals are slightly better at withholding responses to light&tone−. (From data presented in Solomon, 1977, Figure 2.) (B) The cortico-hippocampal model: After 1000 blocks of training, both the intact and HR-lesion models give strong responses to light+ and weaker responses to light&tone−; again, the HR-lesion model learns slightly faster. (C) The S-D model makes the opposite prediction: Although the intact model can learn the task, the lesioned model is unable to discriminate light+ and light&tone− and gives intermediate responses to both. This is because the S-D model assumes that the hippocampus is necessary for cue competition effects. (From data presented in Schmajuk & DiCarlo, 1992, Figure 11.)

our intact and HR-lesioned models show conditioned inhibition, as is seen in figure 6.14B.

The intact S-D model can also produce conditioned inhibition. However, the lesioned S-D model cannot, as is shown in figure 6.14C. This is because the S-D model assumes that hippocampal-region damage disrupts the ability to compute the output error that is needed to learn competing responses to light depending on whether tone is present or absent. Therefore, the lesioned S-D model cannot learn different responses to light and to tone and light and cannot produce conditioned inhibition.[41]

Thus, the data from Solomon and colleagues study of conditioned inhibition are consistent with our cortico-hippocampal model but conflict with the predictions of the S-D model.

Blocking. Another behavioral paradigm in which our cortico-hippocampal model and the S-D model make differing predictions is **blocking.** The basic

Table 6.2 The Blocking Paradigm

Group	Phase 1	Phase 2	Phase 3: Test
Pretrained	Tone → airpuff	Tone&light → airpuff	Tone → ? ⇒ *no blink*
Sit exposure	(Animal sits in experimental chamber)	Tone&light → airpuff	Tone → ? ⇒ *partial blink*

paradigm is schematized in table 6.2. Animals in the Pretrained group receive tone+ training followed by tone&light+ training; when later tested with the light alone, they show no response. By comparison, control animals in the Sit-exposure group, which receive only tone&light training, show at least partial responding to the light stimulus.

Figure 6.15A shows that blocking is unaffected in the HR-lesioned cortico-hippocampal model; by contrast, the lesioned S-D model predicts that the light CS will acquire strong associative strength (figure 6.15B). Unfortunately, the empirical data are ambiguous here; blocking has been reported to be both impaired and spared following hippocampal-region damage in different studies.[42] It is not clear what conclusions to draw from these conflicting results. Perhaps future data will clear up this controversy and provide strong support for one model or the other. Another possibility is that the brain has multiple sites that subserve stimulus competition effects—in the hippocampus and cerebellum and elsewhere—and the particular effects of hippocampal-region damage will depend critically on details of the procedural parameters used in each study. This is especially relevant given that the empirical studies of blocking cited above employed a variety of lesion techniques that produce different degrees of HR damage.

Configural Learning (Negative Patterning). The notion that a task may be either impaired or spared after hippocampal lesion, depending on a variety of procedural details, has implications for a wide range of tasks. One significant task for studies of hippocampal function in recent years has been configural learning, such as the **negative patterning** task. Recall from chapter 3 that this task involves learning to respond to two stimuli (e.g., light+ and tone+) but not to their compound (light&tone−).

Many animal studies have found that hippocampal lesion disrupts the ability to learn negative patterning (e.g., figure 6.16A), but a few studies also showed that negative patterning was spared.[43] Both the cortico-hippocampal

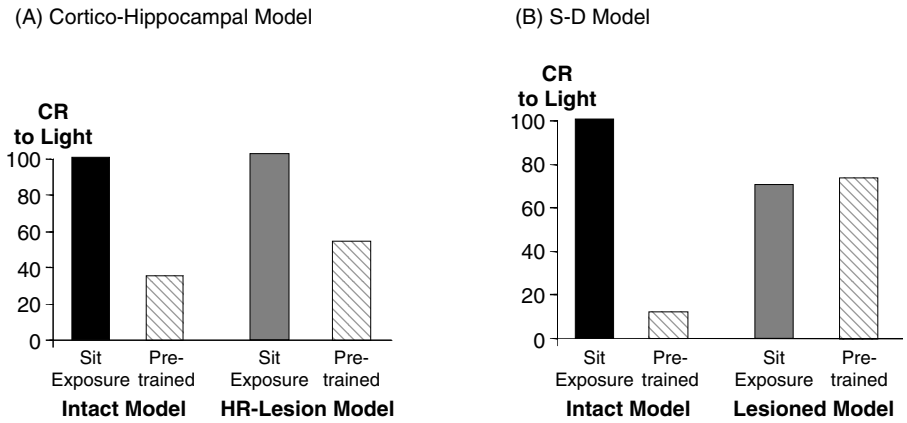


Figure 6.15 Blocking: Tone+ training before tone&light+ training results in a blocking or reduction of associative strength to light. Results are response to light, expressed as a percent of response intact animal or model control condition. (A) The cortico-hippocampal intact and HR-lesion models both show blocking: Responding to light is reduced following tone+ and tone&light+ pretraining relative to a naive condition that just received training to tone&light+. (B) The S-D model predicts that blocking should be eliminated by hippocampal lesion. (Redrawn from data presented in Schmajuk & DiCarlo, 1992, Figure 7.)

and S-D models predict that negative patterning should, in general, be disrupted by hippocampal-region damage, as figures 6.16B and C show. The two computational models make similar predictions here because both assume that the hippocampal region is necessary to form new representations that encode cue configurations (though the two models assume that somewhat different mechanisms underlie this process).

However, both models also assume that hippocampal-region damage will not affect the cue configurations that are already learned and stored outside the hippocampal region (e.g., in cortex). This leads to a subtle prediction of both models: For some configural tasks, preexisting configural representations may exist and suffice to allow the problem to be solved—even without the benefit of hippocampal processing.

Figure 6.17 shows an example from the cortico-hippocampal model:* 100 simulation runs with the intact and HR-lesioned model were trained on the

*Note that the simulations shown in figure 6.17 were not generated from the standard cortico-hippocampal model used elsewhere in chapters 5 and 6 but were instead a version that had fewer internal-layer nodes (four each in the hippocampal region and cortico/cerebellar networks) and also stronger initial weights in the lower layer of the cortico/cerebellar network. This increased the probability that some HR-lesioned networks would reach criterion performance, for demonstration purposes.

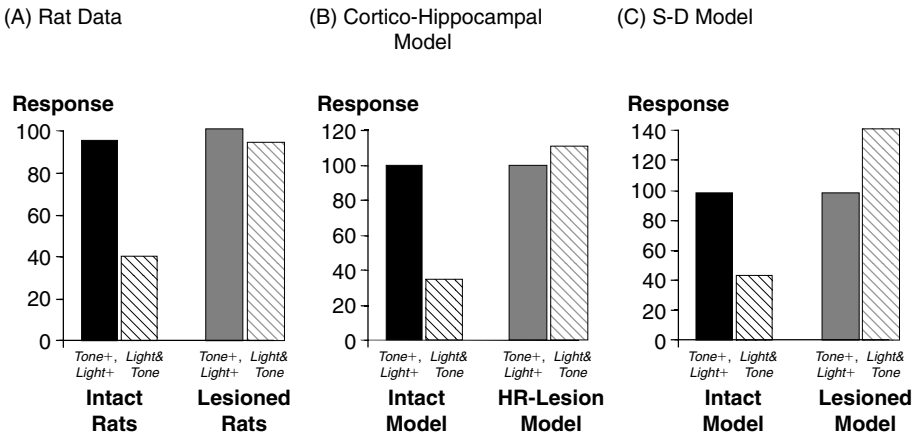


Figure 6.16 Negative patterning involves learning to respond to two cues, light+ and tone+, but to withhold responding to their compound light&tone−. (A) Normal rats can learn to respond to the components but not the compound; hippocampal-lesioned rats are generally found to be unable to withhold responding to the compound light&tone−. Both the cortico-hippocampal model (B) and the S-D model (C) predict that, on average, this and other configural tasks will be greatly disrupted after HR-damage. (A is plotted from data presented in Sutherland & Rudy, 1989, Figure 2. B is plotted from data presented in Schmajuk & DiCarlo, 1992, Figure 14.)

negative patterning task. Solving the task was defined as reaching a criterion performance level, defined as ten consecutive trials on which the simulation generated a CR greater than 0.8 to the components light+ and tone+ but less than 0.2 to the compound light&tone−. By this definition, 80% of the intact simulations did solve the task, reaching criterion performance on discriminating the light+ and tone+ trials from the light&tone− trials. Of the remaining 20%, most simulations still responded somewhat more strongly to the components than to the compound.

The results were very different for the HR-lesion model: 80% of HR-lesion simulations failed to solve the task. Interestingly, the 20% of HR-lesion simulations that did master the task reached criterion performance just as quickly as the intact simulations; in fact, the fastest learners were HR-lesioned, not intact, simulations. This seemingly paradoxical result arises because the intact model is slowed down in learning because it has to construct new stimulus representations to differentiate components light+ and tone+ from the compound light&tone−; in the HR-lesioned model, these representations rarely exist (and hence 80% of the simulations fail), but if by chance the representations do already exist, then learning is very rapid because all that needs to be done to reach criterion performance is to map from these preexisting representations to the correct responses.

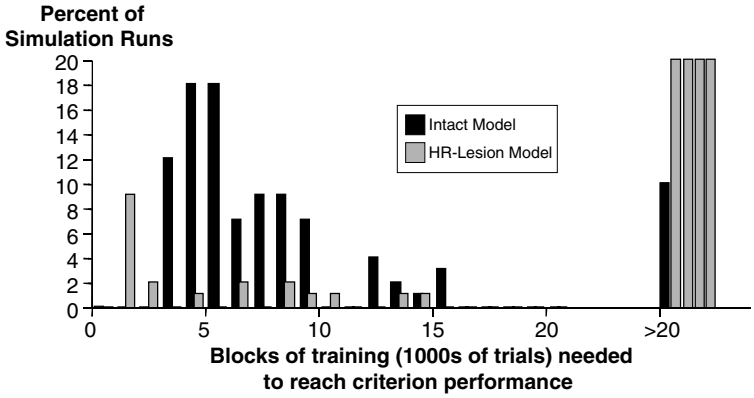


Figure 6.17 Individual performance on negative patterning in the cortico-hippocampal model. The cortico-hippocampal model and S-D model both predict that, while HR lesion will disrupt negative patterning on average, occasionally and seemingly at random an animal's preexisting representations may suffice to solve the problem. Out of 100 simulation runs with both the intact and HR-lesion cortico-hippocampal model, 80% of intact and 20% of HR-lesioned simulations solved the negative patterning problem within 20,000 training blocks. Those HR-lesioned simulations that solved the problem did so as fast as or faster than the intact simulations.

Thus, under certain (possibly rare) conditions, learning in the HR-lesioned model may be faster than in the intact model. Certain kinds of experimental procedures may be especially likely to tap into preexisting, hippocampal-independent systems, in which case configural learning may be reliably faster after hippocampal-region damage. Such results have occasionally been reported in the literature,*[44] lending support to this model prediction.

*Gallagher & Holland, 1992, showed that rats with hippocampal lesions were slightly faster than normal to acquire feature-neutral association (AC+, A−, BC−, B+). Han, Gallagher, & Holland (1998) showed that this effect was strengthened if the time between trials was reduced. Bussey et al., 1998, showed that rats with fornix lesions learned faster than normal on a transverse patterning task (prefer A over B, B over C, and C over A). Eichenbaum & Bunsey, 1995, reported that rats with hippocampal lesions were better than control animals in a paired-associate task (AB+, CD+, AC−, BD−). Each of these tasks embeds configural components. It is very much an open question which precise features of these experiments (including animal species, stimulus modalities, procedural variations, and precise lesion techniques) contribute most to the finding of impaired, spared, or facilitated configural learning after hippocampal-region damage.

6.3 RELATIONSHIP OF MODELS TO QUALITATIVE THEORIES

The previous sections reviewed two computational models that were meant to address information processing in the hippocampal region and how that information might be used by other brain structures during learning. One advantage of computational models is that they provide a reality check: A researcher might propose a mechanism for learning that sounds plausible as a verbal argument, but would not work in practice. If the mechanism can be implemented in a computational model, that suggests that the mechanism would indeed operate as expected. A computational model does not *prove* that the brain operates in a certain way, but if the model successfully accounts for a large portion of existing data and makes predictions that are later confirmed by experimental testing, then that suggests that the model is on the right track.

A second use of computational models is that they can often provide possible explanations for issues that were not obvious before. For example, the cortico-hippocampal and S-D models provide one interpretation for the paradoxical finding that, although configural learning may often be devastated following hippocampal-region damage, in some rare cases it is spared or even facilitated. *Thus, it may not be particularly useful to attempt to dichotomize tasks according to whether they can or cannot survive hippocampal-region damage. The computational-modeling approach suggests that it is more useful to consider what kinds of information the hippocampal region normally processes and which tasks may be expected generally to depend on this information.*

Other researchers have also taken information-processing approaches to understanding the hippocampal region and have developed qualitative theories about the hippocampal region's role that are often very useful as heuristics to predict behavior in the lesioned animal or human. In many cases, these qualitative theories are consistent with the computational models described above. This section reviews several prominent qualitative theories of hippocampal-region function, and notes how they relate to the computational models described in sections 6.1 and 6.2.

Stimulus Configuration

Several prominent theories of hippocampal-region function have assumed that the hippocampal region is involved in **stimulus configuration** (or **"chunking"**), whereby a set of co-occurring stimuli come to be treated as a unary whole (or configuration) that can accrue associations.[45] For example, the negative patterning task requires subjects to learn to respond to two

stimuli (light+ and tone+) but not their compound (light&tone−). This is easily solved if it is assumed that light and tone can enter into direct excitatory associations with the US but that the compound light&tone is a separate entity that can enter into direct inhibitory associations with the US. In fact, early studies demonstrated that such configural tasks were especially sensitive to hippocampal-region damage.[46] However, later studies suggested that hippocampal-lesioned animals could indeed solve some configural problems.[47] While configuration may be especially sensitive to hippocampal-region damage, it is clearly not universally abolished by such damage.

An alternative interpretation of configuration is embodied by the cortico-hippocampal model. The model assumes that the hippocampal region forms new stimulus representations that may compress (or chunk or configure) co-occurring stimuli. This will facilitate learning of such tasks as negative patterning that depend on stimulus configuration. In the lesioned model, the cortico/cerebellar network is left with a set of fixed lower-layer weights, which do perform a recoding of stimulus inputs. Depending on the initial state of these weights, there is always some probability that a random configuration may be encoded by those weights. In such a case, a configural task may well be solved by the lesioned model. Thus, the cortico-hippocampal model expects only a general tendency toward a lesion deficit in configural learning, not an absolute deficit. This interpretation is consistent with the finding in a few studies that, although lesioned animals are generally impaired at a configural task, they may occasionally solve a configural task as quickly as—or faster than—control animals.[48] It is also consistent with the observation, in many studies, of wide individual variance in how animals solve configural tasks. For example, near the end of a study of negative patterning in normal rabbits, most animals were reliably giving eyeblink CRs to the components light+ and tone+ but not to the configuration light&tone−.[49] However, about one-fourth of the rabbits were showing a different pattern: responding strongly to one component and weakly to the configuration—but also failing to respond to at least one of the components. An additional rabbit was showing the opposite failure: responding reliably to both components and also responding strongly to the compound.[50] Thus, as in the model in figure 6.17, some individual rabbits learn quickly and some learn slowly or not at all. The variation can be even more pronounced in lesioned animals.

Thus, the computational models implement much of the spirit of the configural theory while also demonstrating how and why exceptions to the rule might exist.

Contextual Learning

A related class of theory implicates the hippocampal region in contextual processing. Specifically, the hippocampus has been proposed as the source for *contextual tags* to memories, which identify the spatial and temporal settings in which events occur.[51] **Context** is usually defined as the set of background cues that are present during a learning experience but distinct from the experimentally manipulated CS and US; example contextual stimuli include visual features of the room, background noises, temperature, and time of day. Contextual processing is often disrupted after hippocampal-region damage. For example, humans with medial temporal lobe amnesia can often acquire new information but not recall the spatial and temporal context in which it occurred.[52] Conversely, the nondeclarative (or incremental or procedural) learning that is often preserved in amnesia tends to be acquired slowly, over many trials, and therefore is less strongly associated with any particular context.

Chapter 7 will consider the role of the hippocampal region in contextual processing in more detail; for now, it is sufficient to note that *the cortico-hippocampal model assumes that all co-occurring stimuli (CS, US, and context) influence the development of hippocampal-mediated stimulus representations*. Thus, if CS-US learning occurs in one context, the representation of CS will include contextual information. In contrast, the lesioned model does not develop new stimulus representations but only forms direct CS-US associations; in most cases, the context will not be sufficiently predictive of the US to enter into any associations. Thus, in many cases, there are differences between the context-sensitivity of the intact and lesioned model, and these differences parallel the behavior of intact and lesioned animals.

Stimulus Selection

Some early theories of hippocampal function viewed the hippocampal region as an attentional control mechanism, responsible for reducing attention to stimuli that are not significant, are not correlated with reinforcement, or are irrelevant with respect to predicting reinforcement.[53] These theories are concerned with **stimulus selection:** how individual stimuli are "tuned in" or "tuned out" of attention.

Traditionally, psychological theories of stimulus selection have fallen into two broad classes: **reinforcement modulation theories** and **sensory modulation theories.** Reinforcement modulation theories consider the effectiveness of the reinforcer (e.g., the US) to be modulated by the degree to which

the US is unexpected, given all the cues (e.g., the CS) present. So a reinforcer that is predicted by a CS would not support new learning, while one that is unexpected would support new learning. The Rescorla-Wagner model described in chapter 3 is an example of a reinforcement modulation theory.

In contrast, sensory modulation theories of stimulus selection focus on the ability of CSs to enter into new associations, on the basis of how much they add to the overall ability to predict reinforcement.[54] Thus, a CS that predicted an otherwise unpredictable US would be very likely to enter into associations; a CS that did not add to the ability to predict the US would not enter into new associations. Both reinforcement and sensory modulation theories can account for some—but not all—learning behaviors.

Interestingly, a few behaviors, such as blocking, can be explained in terms of either (or both) approaches: At the end of phase 1, the trained CS (e.g. tone+) perfectly predicts the US. When the second light+ CS is added, reinforcement modulation theories expect that the well-predicted US should not enter into new associations with the light. At the same time, sensory modulation theories expect that the light, which adds no new information, should not enter into new associations with the US. Thus, both sensory modulation and reinforcement modulation may normally contribute to a strong blocking effect.

The cortico-hippocampal model and the S-D model of hippocampal-region function incorporate both sensory and reinforcement modulation. The cortico-hippocampal model assumes that the cortico/cerebellar regions (which are capable of error-correction) can perform reinforcement modulation, while the hippocampal region is needed for sensory modulation. Thus, our model predicts that some—but not all—stimulus selection effects will be disrupted after hippocampal-region damage. Specifically, reinforcement modulation should survive HR lesion, while sensory modulation should not.

More generally, because stimulus selection depends on two substrates (cortico/cerebellar and hippocampal), our model expects that many behaviors that reflect stimulus selection may be reduced but not eliminated by HR lesion, because removing one of two sources of stimulus selection may leave the other intact. This may explain why some stimulus selection behaviors, such as blocking, sometimes survive hippocampal lesion and sometimes are eliminated.

Intermediate-Term and Working Memory

The fact that HR-lesioned monkeys are impaired at delayed nonmatch to sample—but only if there is a long enough delay between sample and choice—suggests that the hippocampal region has a role in maintaining

information over the course of a few minutes. This kind of memory is often called **intermediate-term memory** (as distinct from the short-term memory that we use to remember a telephone number by constant rehearsal or long-term memory, which can last years). A related concept is **working memory:** intermediate-term memories that contain information relevant to the current task at hand. For example, in the monkey delayed nonmatch to sample (DNMS) paradigm, working memory allows the monkey to remember the sample item across a short delay; however, on each new trial, the sample item is different, and working memory updates to reflect this. A working-memory task in rats might involve a maze in which there is food in a number of locations at the start of each trial; to obtain all the food in the shortest possible time, the rat must use working memory to remember which locations it has visited (and depleted) so far on the current trial. Performance on this kind of task is disrupted by hippocampal-region damage.[55] Trace conditioning also requires the ability to maintain CS information during the short interval until the US arrives and is likewise disrupted by hippocampal-region damage.

Several researchers have suggested that a critical component of these tasks is the ability to represent information over short periods of time, and some have suggested that the hippocampus functions as a buffer, holding critical bits of information for a short time.[56] In fact, there are some neurophysiological data suggesting that neuronal activity in the hippocampus may temporarily encode recent stimuli.[57] More recently, researchers have suggested that the hippocampus is critical when the task has a temporal discontinguity, meaning that the items to be associated do not overlap in time.[58]

Howard Eichenbaum, Tim Otto, and Neal Cohen have suggested that the intermediate-term memory buffer can be specifically localized within the **parahippocampal region,** including entorhinal, perirhinal, and parahippocampal cortices.[59] One function of this buffer would be to compress or "fuse" individual stimuli into compound percepts.[60] We will return to the possible selective role of entorhinal cortex and other parahippocampal structures in chapter 9; for now, we note that the stimulus fusion that Eichenbaum and colleagues attribute to the entorhinal cortex is perfectly consistent with the stimulus compression that the cortico-hippocampal model views as one aspect of hippocampal-region function.[61] The hippocampal region may be involved in constructing new representations that compress together stimuli that co-occur or are similar in meaning; this compression could apply equally to stimuli that are superficially dissimilar, and to those that are separated in time or in space.[62]

Cognitive Mapping

Perhaps the most devastating effect of hippocampal-region damage in rats is on spatial learning. The strong spatial impairment in hippocampal-lesioned rats has led to theories suggesting that the hippocampus (or hippocampal region) is specialized as a spatial mapping system.[63] Partial support for these theories comes from neurophysiological studies showing that individual cells in hippocampal subfields CA3 and CA1 respond preferentially when the animal is in a particular region of space. A variety of computational models have shown that the known anatomy and physiology of these subfields is sufficient to give rise to place field behavior and allow various kinds of spatial learning.[64]

Perhaps the primary problem with the simplest spatial theories is that hippocampal-region damage can result in a broad range of deficits that have no obvious spatial component; examples include conditioning tasks such as disrupted sensory preconditioning in HR-lesioned animals and disrupted recognition in humans with medial temporal damage (refer to chapter 1). The spatial theory has since been extended to assume that the hippocampus is only disproportionately, but not exclusively, involved in spatial processing[65] or that the hippocampal region is involved in **cognitive mapping,** which ties spatial as well as contextual, semantic, and other information into unified memories.[66] This, then, would lead to the impairment in episodic learning observed in human amnesia.

An alternative interpretation of the lesion impairment in spatial learning suggests that the hippocampus is not specialized for spatial learning per se, but rather that a "place" is simply a configuration of local views of space.[67] Thus, since configural processes seem to be especially sensitive to hippocampal-region damage, spatial learning is also susceptible to hippocampal-region impairment. This interpretation is consistent with the fact that some hippocampal cells that show spatially determined responses during a spatial task can have other behavioral correlates during a nonspatial task.[68]

"Flexible" Memory

In contrast to theories that implicate the hippocampal region specifically in one kind of memory—be it configural, contextual, spatial, or even declarative—some researchers have suggested that the hippocampal region of an intact animal is involved in learning even the most elementary associations. For example, in the cortico-hippocampal model, the hippocampal region is

not *necessary* for learning a CS-US association; however, in the intact model, the hippocampal region constructs new representations even during such a simple task and thereby influences the way in which CS-US associations are formed. This may not be evident during learning, and so there may be little or no difference in the CR development of an intact or hippocampal region-lesioned animal (or model). But the hippocampal-mediated representations will become very evident if the animal (or model) is later challenged to use that learning in a generalization or transfer task.

Eichenbaum, Cohen, and colleagues have advanced a qualitatively similar view.[69] They propose that the hippocampus is involved in forming stimulus representations that are sensitive to the relationships between stimuli.[70] Hippocampal-independent memories are assumed to be inflexible, in the sense that they can be accessed only through reactivation of the original stimuli and situations in which the learning took place. For example, rats can be trained to prefer odor A+ over odor B− and to prefer odor C+ over D−. Later, when presented with odors A and D, normal rats reliably choose A but hippocampal-lesioned rats may choose randomly—as if they had never been exposed to either odor before.[71] In chapter 8, we will discuss this kind of task in more detail and show how the cortico-hippocampal model can be extended to address some of its features. For now, though, we simply note a surface similarity between Eichenbaum and Cohen's theory and the cortico-hippocampal model. A central feature of Eichenbaum and Cohen's flexible learning is the creation of stimulus representations that emphasize the relationships between stimuli. This is consistent with the cortico-hippocampal model, in which hippocampal-region representations emphasize predictive features and deemphasize irrelevant features; such representations may be used flexibly in new contexts whose irrelevant features differ from those of the learning context.

6.4 IMPLICATIONS FOR HUMAN MEMORY AND MEMORY DISORDERS

The first and most important implication of the computational models discussed in this chapter is that hippocampal-region damage does not indiscriminately abolish the ability to learn. Animals with HR lesion can acquire some kinds of new information, including simple acquisition of conditioned eyeblink responses and other stimulus-response associations.

It is now well established that humans with anterograde amnesia from medial temporal (including hippocampal-region) damage can also learn motor-reflex responses.[72] Similarly, by focusing on what amnesic subjects can do rather than on their impairments, it may be possible to accomplish

significant and subtle learning in these individuals. Amnesic subjects are able to acquire and retain new motor skills such as mirror-tracing and cognitive skills such as grammatical rules if the learning takes place incrementally, over many trials, and does not require episodic memory of individual learning sessions.[73] Similarly, amnesic subjects can learn to categorize objects on a computer screen into arbitrary classes by repeated exposure to members of the categories.[74] Amnesic subjects have even been taught rudimentary computer programming skills using methods that take advantage of spared learning abilities.[75]

The animal and model data also suggest that some kinds of conditioning should be disrupted by HR lesion. For example, discrimination reversal may be greatly slowed in relation to controls, although HR-lesioned animals may master the reversal given enough trials. One question is whether humans with medial temporal (HR) damage will show a similar pattern of impaired and spared learning. At this point, it is still very much an open question whether amnesic individuals can reverse a learned discrimination as well as control subjects do.[76]

In other kinds of paradigm, in which control animals are slowed by prior exposure, HR-lesioned animals can outperform control animals. For example, controls but not HR-lesioned animals are slow to learn a CS-US association following uncorrelated exposure **(learned irrelevance)** to the CS and US. Recently, we developed in our laboratory a computerized task that embeds some features of learned irrelevance into a video game task.[77] The task involved learning that some screen events (like CSs) predict other screen events (like USs). Subjects were seated at a computer screen and told that they would see a magician trying to make a rabbit appear under his hat (figure 6.18). On each trial, the subjects were to guess whether or not the magician succeeded. The appearance of the rabbit was conceptually analogous to the to-be-predicted US, and the subjects' predictions were equivalent to an anticipatory CR.

Subjects in our study were divided into two groups: Exposed and Nonexposed. In phase 1, for all subjects, the appearance of the rabbit US was contingent on a particular "magic word" in the magician's cartoon word balloon. For subjects in the Nonexposed group, the cartoon balloon was always uncolored (gray). For subjects in the Exposed group, the cartoon balloon was colored red or green, and the color was not correlated with the rabbit US. Later, in phase 2, the balloon color perfectly predicted the rabbit US. Thus, balloon color was the CS that predicted US arrival. Subjects who had previously been exposed to this color CS, uncorrelated with the US, were slower to learn the association between CS and US in phase 2 than were

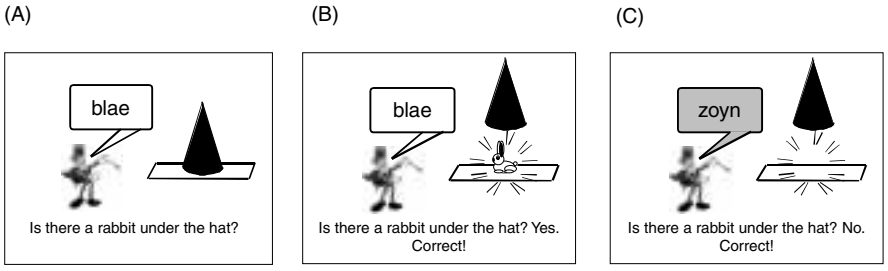


Figure 6.18 A computerized task that embeds some aspects of learned irrelevance. Subjects watch a computer screen that shows a magician trying to produce a rabbit under his hat (A). Conceptually, the appearance of the rabbit is the to-be-predicted event (US). Subjects guess whether the rabbit will appear on each trial. The hat is then raised to show whether the rabbit is present (B, C), and corrective feedback is given. In phase 1, the rabbit is predicted by a particular magic word in the magician's cartoon word balloon. In phase 2, the rabbit is predicted by a particular color (red or green) in the word balloon. (Adapted from Myers, McGlinchey-Berroth, et al., 2000, Figure 1.)

subjects who had not been exposed. Thus, normal subjects showed learned irrelevance.[78]

Recently, we tested a group of individuals with amnesia resulting from medial temporal (HR) damage and a group of matched control subjects on this task.[79] In phase 1 (in which the "magic word" predicted the rabbit), amnesic and control subjects all learned quickly, regardless of exposure condition (figure 6.19A). In phase 2, the control subjects showed a strong learned irrelevance effect: The Exposed group learned more slowly than the Nonexposed group. However, among amnesic subjects, the learned irrelevance effect was eliminated: Exposed and Nonexposed groups learned at the same rate (figure 6.19B).

One implication of these findings in animals and humans is that *under certain conditions that slow CS-US learning in normal subjects, the amnesic subjects learn more quickly than controls!* Of course, the normal subjects learn more slowly because they are learning more: They are learning about environmental regularities during exposure, and this same learning is what disrupts later association.

A second implication of these findings is that *although both controls and amnesic subjects appear to learn similarly in phase 1, they are actually learning differently*. This difference shows up during phase 2, when subjects are challenged to apply their learning in new ways. Thus, transfer tasks may be a more useful way to demonstrate differences between these groups than the initial learning. This in turn has potential implications for diagnosing syndromes,

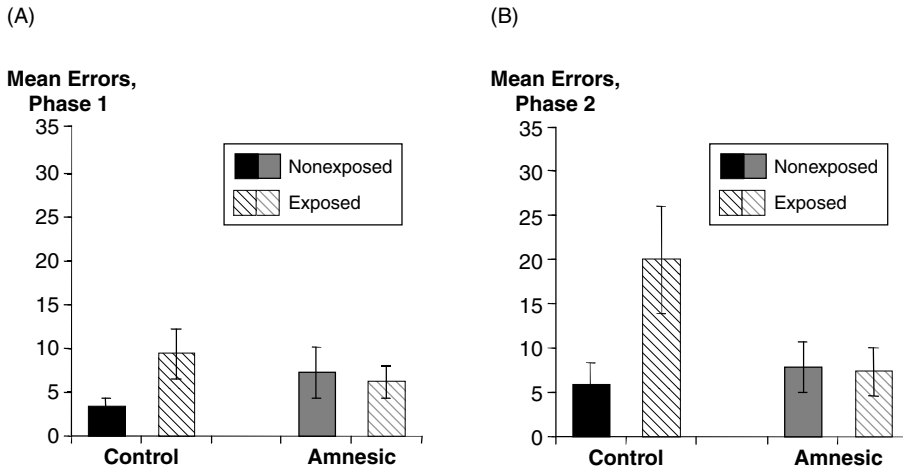


Figure 6.19 Results from the learned irrelevance paradigm shown in Figure 6.19. (A) In phase 1, subjects learned to predict the rabbit US on the basis of a neutral cue (a particular magic word). Subjects in the exposed condition were also given uncorrelated exposure to a color CS; subjects in the nonexposed condition never saw this CS in phase 1. Among normal subjects, there was no significant difference in total errors between the two conditions; amnesic subjects in both conditions also performed the same as controls. (B) In phase 2, the color CS predicted the rabbit US. Normal subjects who had been exposed to the CS in phase 1 took longer to learn this CS-US association than did nonexposed subjects. Thus, normal subjects showed learned irrelevance. By contrast, amnesic subjects did not show learned irrelevance: Exposed and nonexposed amnesic subjects learned the phase 2 task at the same speed—and much faster than exposed control subjects. (Adapted from Myers, McGlinchey-Berroth, et al., 2000, Figure 2.)

such as Alzheimer's disease, that involve hippocampal-region damage. We will consider the application of related human experiments to the prediction of the early stages of Alzheimer's in chapter 9.

SUMMARY

• Hippocampal-region damage in network models is simulated by disabling a hippocampal-region module and observing the behavior of the remaining modules.

• Gluck and Myers's cortico-hippocampal theory argues that the hippocampal region plays a crucial role in the recoding or rerepresentation of stimulus representations during learning. Specifically, if two stimuli co-occur or make similar predictions about future reinforcement, their representations will be compressed to increase generalization between the stimuli. Conversely, if two stimuli never co-occur and make different predictions about future

reinforcement, their representations will be differentiated to decrease generalization between the stimuli.

- The cortico-hippocampal model assumes that the representations developed in the hippocampal region are eventually adopted by other long-term storage sites in cortex and cerebellum.
- The hippocampal region may not be strictly *necessary* for some simple kinds of learning; but when it is present, it normally contributes to *all* learning.
- The cortico-hippocampal model has limitations in accounting for data on extinction and response timing.
- Taken together, the neurophysiological evidence that is currently available is remarkably consistent with the implications of our cortico-hippocampal model, suggesting that hippocampal neuronal representations can and do change to reflect associations between stimuli and rewards, much like the internal-layer nodes in a predictive autoencoder.
- The Schmajuk-DiCarlo (S-D) model assumes that CS information reaches the cerebellum via two routes: a direct path and an indirect path involving association cortex. Specifically, the hippocampus in this model is presumed to calculate the predicted US, while other brain areas compare this predicted US against the actual US and calculate the total error. This error signal is then used to guide learning. In addition, cerebellar units can update weights directly on the basis of the error signal, whereas the cortical units require specialized error signals broadcast by the hippocampus.
- It may not be particularly useful to attempt to dichotomize tasks according to whether they can or cannot survive hippocampal-region damage. Computational modeling suggests that it is more useful to consider what kinds of information the hippocampal region normally processes and which tasks may be expected generally to depend on this information.
- In many cases, qualitative theories of the hippocampal region are very consistent with a subset of the implications of computational models when the models are applied to specific domains or tasks.
- Because selective attention depends on both cortico/cerebellar and hippocampal substrates in the cortico-hippocampal model, the model expects that many behaviors that reflect selective attention may be reduced but not eliminated by HR lesion.
- In studies of associative learning in cognitive analogs of conditioning tasks, both control and amnesic subjects appear to learn similarly during an initial phase of training, but transfer task performance suggests that they are actually using different strategies to learn.

APPENDIX 6.1 SIMULATION DETAILS

The cortico-hippocampal model of section 6.1 was described at a very general level, without reference to mathematical details such as number of nodes or learning rates. In fact, over the years, we have implemented the model with a wide number of parameter choices; the qualitative behavior of the model is usually independent of these choices, although the absolute speed or accuracy of learning may vary. All of the cortico-hippocampal model simulations presented in section 6.1 were based on a single implementation (the same one used in Myers & Gluck, 1994), and each figure represents the average of ten simulation runs with that implementation, except as otherwise noted in the text.

The external inputs consisted of four CSs and fourteen contextual stimuli, each of which could be either present or absent. The contextual stimuli were initialized randomly but thereafter held constant for the remainder of the experiment. (A different set of randomly initialized contextual stimuli was used if there was to be a contextual shift in the experiment.) At any given moment, one or more CSs could be present along with the US. A **trial** consisted of one presentation of each stimulus combination to be trained, interspersed with context-alone trials in a 1:19 ratio. Thus, for discrimination learning (CS A predicts the US, but CS B does not), one trial would include a presentation of A (with the US), nineteen context-alone presentations, a presentation of B (with no US), and nineteen more context-alone presentations.

The hippocampal-region network contained eighteen input nodes, ten internal-layer nodes, and nineteen output nodes. The output nodes learned to reconstruct the eighteen inputs as well as predicting whether the US was present. The network was trained by error backpropagation, as described in MathBox 4.1 and Rumelhart, Hinton, and Williams (1986). The learning rate was set at 0.05 when the US was present and 0.005 otherwise; the momentum was set at 0.9.*

The cortico/cerebellar network also contained eighteen input nodes, along with sixty internal-layer nodes and one output node. The activity of the output node was interpreted as the strength or probability of a CR in response to the current inputs. The upper layer of weights was trained according to the Widrow-Hoff rule (MathBox 3.1); desired output was the same as the US. The

*Note that there is a wide range of parameters for the model that would produce similar rates of learning; for example, setting the learning rates to be equivalent on US-present and US-absent trials but greatly increasing the ratio of context-alone versus CS+ trials would produce generally similar, though not necessarily identical, behavioral properties.

learning rate was set at 0.5 when the US was present and 0.05 otherwise. For each internal-layer node c, the desired output was defined as the sum over all hippocampal-region network internal-layer nodes h of the activity of h times the weight from h to c. These weights were initialized according to the random distribution $U(-0.3. \ . \ .+0.3)$ and were held constant throughout the experiment.

All nodes in the system used a sigmoidal output function as defined in MathBox 4.1, Equation 4.3. All weights and biases in the hippocampal-region network were initialized according to the random distribution $U(-0.3. \ . \ . \ +0.3)$. Upper-layer weights and output bias in the cortico/cerebellar network were initialized in the same way. Lower-layer weights and internal-layer node biases in the cortico/cerebellar network were initialized according to the random distribution $U(-15. \ . \ . \ +15)$. This, in combination with the large number of internal-layer nodes in the cortico/cerebellar network, maximized the chance of useful representations existing in the HR-lesioned model.

Before any training, the entire system was initialized with 500 context-alone presentations, simulating the time spent acclimatizing an animal to the experimental chamber before any conditioning begins. The model was assumed to have "learned" a task when it reached **criterion** performance, defined as ten consecutive trials correctly generating a CR of at least 0.8 whenever the US was present and at most 0.2 when the US was absent.