HIPPOCAMPUS, VOL. 3, NO. 4, PAGES 491–516, OCTOBER 1993

Hippocampal Mediation of Stimulus Representation: A Computational Theory

Mark A. Gluck and Catherine E. Myers

Center for Molecular and Behavioral Neuroscience, Rutgers University, Newark, New Jersey, U.S.A.

ABSTRACT

The authors propose a computational theory of the hippocampal region's function in mediating stimulus representations. The theory assumes that the hippocampal region develops new stimulus representations that enhance the discriminability of differentially predictive cues while compressing the representation of redundant cues. Other brain regions, including cerebral and cerebellar cortices, are presumed to use these hippocampal representations to recode their own stimulus representations. In the absence of an intact hippocampal region, the theory implies that other brain regions will attempt to learn associations using previously established fixed representations. Instantiated as a connectionist network model, the theory provides a simple and unified interpretation of the functional role of the hippocampal region in a wide range of conditioning paradigms, including stimulus discrimination, reversal learning, stimulus generalization, latent inhibition, sensory preconditioning, and contextual sensitivity. The theory makes novel predictions regarding the effects of hippocampal lesions on easy-hard transfer and compound preexposure. Several prior qualitative characterizations of hippocampal function—including stimulus selection, chunking, cue configuration, and contextual coding—are identified as task-specific special cases derivable from this more general theory. The theory suggests that a profitable direction for future empirical and theoretical research will be the study of learning tasks in which both intact and lesioned animals exhibit similar initial learning behaviors but differ on subsequent transfer and generalization tasks.

Key words: neurocomputation, computational models, representation, declarative memory, conditioning, connectionist models

Although the hippocampus and adjacent cortical regions are generally acknowledged to play a fundamental role in learning and memory, little consensus has emerged as to the precise specification of this role. Lesion data from humans (Scoville and Milner, 1957; Squire, 1987) and animals (Mishkin, 1982; Squire and Zola-Morgan, 1983) suggest that some learning tasks may not be solvable without an intact hippocampal region. One approach to understanding hippocampal function has thus been to characterize the class of learning and memory tasks that require an intact hippocampal region. Squire and colleagues have emphasized the critical role of this brain region for the formation of explicit declarative memory in humans (Squire, 1987; Squire and Frambach, 1990). Studies of lower (nonprimate) mammals have focused on place learning and spatial navigation as tasks that require an intact hippocampal region (O'Keefe and Nadel, 1978; Morris et al., 1982; McNaughton and Nadel, 1990).

Another approach to theories of hippocampal region function has been to identify a putative information processing role for the hippocampal region. In this way, theorists seek to derive a broad range of task-specific deficits in lesioned animals from some underlying conceptualization of hippocampal function. Two broad classes of hippocampal functions have been considered. The first class includes temporal processing deficits that occur following hippocampal lesions: these include impairments with sequence learning (e.g., Buszaki, 1989) and response timing (Akase et al., 1989; Moyer et al., 1990). A second class of hippocampal-dependent behaviors appear to concern aspects of stimulus representation; these focus on hippocampal-lesion deficits in learning complex tasks or in the flexible use of learned information in novel situations. In the analyses to follow, we will focus exclusively on the latter, representational, processes that can be seen in trial-level analyses of conditioning. For this reason, this report will not address the many known temporal aspects of hippocampal function.

Several representational theories of hippocampal region function have been proposed. Early theorists viewed the hippocampus as an attentional control mechanism that alters stimulus selection through a process of inhibiting attentional

Correspondence and reprint requests to Mark A. Gluck, Center for Molecular and Behavioral Neuroscience, Rutgers University, Newark, NJ 07102 U.S.A.

responses to irrelevant cues (Douglas and Pribram, 1966; Schmajuk and Moore, 1985). Wickelgren (1979) suggested that the hippocampus participates in a process whereby component features within a stimulus pattern are recognized as co-occurring elements and thus come to be treated as a unary whole or "chunk." More recently, Wickelgren's chunking idea has been extended and elaborated by Sutherland and Rudy (1989) who propose that the hippocampus provides the neural basis for the acquisition and storage of configural cues. Still others have suggested that the hippocampal region provides a "contextual tag" for associative learning (Hirsh, 1974; Nadel and Willner, 1980; Winocur et al., 1987; Penick and Solomon, 1991). Eichenbaum and colleagues have emphasized the representational role of hippocampal function, especially in the flexible use of conjunctive and relational associations in novel situations (Eichenbaum and Buckingham, 1991; Eichenbaum et al., 1992). McNaughton and colleagues have argued that similar representational constraints might underlie hippocampal involvement in spatial tasks (McNaughton and Nadel, 1990; McNaughton et al., 1992). These previous qualitative views of hippocampal function will be reviewed in more detail and compared to our current computational theory later in this report.

We have taken a slightly different approach than previous theorists by starting with a computational theory of discrimination learning in intact animals and then seeking to identify a subcomponent of this theory that depends on the hippocampal region. This leads us to a computational theory of the functional role of the hippocampal region in mediating stimulus representation. The theory assumes that the hippocampal region develops new stimulus representations that enhance the discriminability of differentially predictive cues while compressing the representation of redundant cues. Other brain regions, including cerebral and cerebellar cortices, are presumed to use these hippocampal representations to recode their own stimulus representations. In the absence of an intact hippocampal region, the theory expects that other brain regions can still acquire some associative learning tasks using previously established fixed representations. However, this a-hippocampal learning is expected to show consistent and predictable differences from normal learning, especially inappropriate generalization to novel task demands.[1]

Once these representations are fully acquired by the cortical regions responsible for long-term memory and response generation, the memories may survive subsequent hippocampal damage, although there may be some consolidation period during which cortical memories depend on an intact hippocampal region. We will not address this putative consolidation process in this article; instead, we focus on comparisons of intact animals with animals whose hippocampal regions have been lesioned (or otherwise inactivated) prior to training. We note here only that our basic theory is amenable to some unspecified process of consolidation wherein the hippocampal region would effect the acquisition of new memories but not their long-term retention.

To develop our theory, we use connectionist network models as a formal framework—or language—for characterizing theories of associative learning. Within this framework, we develop and test a trial-level connectionist theory of cortico-hippocampal interaction in classical conditioning. We argue that the theory provides a simple and unified interpretation of the functional role of the hippocampal region in a wide range of conditioning paradigms, including stimulus discrimination, reversal learning, stimulus generalization, latent inhibition, sensory preconditioning, and contextual sensitivity. The theory also makes novel predictions regarding the effects of hippocampal lesions on several additional training paradigms.

We have focused on classical conditioning for two reasons. First, there exist extensive behavioral studies of both intact and hippocampal-lesioned animals for this paradigm. Second, classical conditioning has received extensive analyses at the behavioral level, resulting in powerful and elegant computational models that characterize the behavioral properties of this form of learning (Sutherland and Mackintosh, 1971; Rescorla and Wagner, 1972; Pearce and Hall, 1980; Sutton and Barto, 1981; Mackintosh, 1983). These behavioral models provide an important starting point for understanding and characterizing what is missing or different in hippocampal lesioned animals. Although our model of the intact animal can be evaluated as a psychological (i.e., behavioral) theory of discrimination learning, our primary goal in this paper is to evaluate the theory as a psychobiological characterization of the functional role of the hippocampal region. Later, however, we will briefly note how this theory can be viewed as a reconciliation of two disparate theoretical approaches found in the psychological literature for dealing with the problem of selective attention in discrimination learning (e.g., Sutherland and Mackintosh, 1971; Rescorla and Wagner, 1972).

CORTICO-HIPPOCAMPAL THEORY

In our cortico-hippocampal theory, novel stimulus representations constructed by the hippocampus are constrained by two biases. These provide additional constraints—above and beyond the constraints imposed by the training task—that determine how stimuli will be represented. The idea of constraints on representations can most easily be understood within the framework of a connectionist network. Within a multilayer connectionist network there may be many (possibly infinite) combinations of weights that can adequately solve a given training task. The training task, in this case, is said to underconstrain the solution. The imposition of additional biases by the hippocampal region can be viewed as constraining (and, hence, narrowing) the set of allowable solutions. We will briefly introduce these two representational biases. Later, they will be described in more detail when we present a specific connectionist network model that constructs stimulus representations constrained by these biases.

The first hippocampal-dependent constraint on representations is a bias to enhance the discriminability of stimuli that predict different outcomes. This results in an increase in the

[1] The computational function we propose here depends critically on intact hippocampal region. It may be sited in one or more of the structures comprising the hippocampal region, or may be distributed throughout them. It is also possible that the function is actually sited externally, for example in the cortices itself, and merely depends in some other way upon intact hippocampal region. Without disavowing either possibility, we refer for simplicity in this paper to the hippocampal region as the site of execution of the proposed function.

resources allocated to represent stimulus features that have a predictive value. We refer to this constraint as a bias for *predictive differentiation*. The second hippocampal-dependent constraint is a bias to efficiently and compactly encode stimuli using minimal resources. This has the effect of combining or clustering correlated (and hence redundant) stimulus features. We refer to this constraint as a bias for *redundancy compression*. The connectionist network model of cortico-hippocampal processing described below illustrates how many associative learning behaviors (and their disruption in hippocampal impaired animals) can be understood as resulting from one or the other, or both, of these representational biases.

Connectionist network models

Connectionist networks are a mathematical formalism for describing associative systems that propagate activation in parallel among many elements (Rumelhart and McClelland, 1986). This formalism can be viewed as a framework for modeling computational theories of associative learning. While these models are similar to neuronal networks in the brain, they do not directly map onto physiological descriptions of circuit-level processing. Nevertheless, when viewed as psychobiological theories of associative learning, these models can illuminate how information processing might be distributed among distinct anatomical brain regions.

Figure 1 illustrates a generic multilayer connectionist network for associative learning. Stimulus information enters at the bottom through the stimulus (input) nodes. Activation of these input nodes propagates activation through weighted connections to the middle layer of nodes. Activation in this middle layer is a function of the weighted sum of incoming activations from the input layer. The middle layer can be viewed as the network's *internal representation* of the stimulus pattern. This internal representation can be contrasted with the (external) stimulus representation on the input nodes.

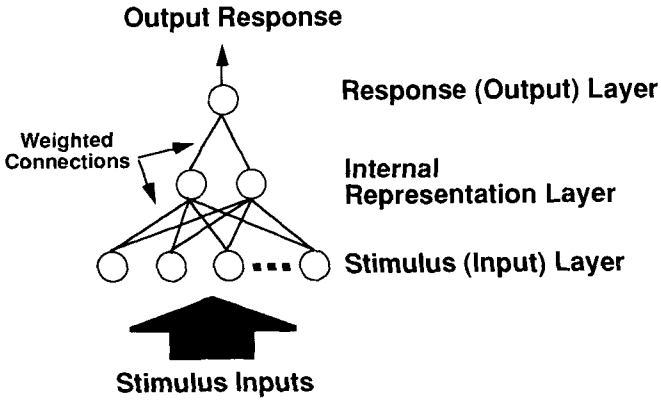


Fig. 1. A generic connectionist network for associative learning. Stimulus information enters through the input nodes and propagates through weighted connections to the internal layer nodes. The activation of these internal layer nodes constitutes the network's internal representation. These activations propagate through a second layer of weighted connections to the output layer. Output layer activations can be interpreted as the network's response.

The internal pattern of activation is itself propagated up through the next layer of weights to the output (response) layer. Activation in the response layer is a function of the weighted sum of activations in the previous layer.

Within a general connectionist framework, models of specific psychological processes or biological regions are formulated by specifying the connectivity (or architecture) of the network, the learning rules for updating associative weights, and the mapping from external sensory stimuli to network input and from network output to observable response behaviors (see Gluck and Bower, 1988a, 1988b, 1990).

Stimulus representation

Animal's response to a sensory stimulus cue, such as a tone, can be modeled as follows. The presence of a tone, along with assorted contextual cues, is represented by a pattern of activation on the stimulus (input) layer (Fig. 2A). The unspecified contextual cues might include the experimental apparatus or other environmental cues and are characterized in the model by a randomly chosen set of activations. Input activation is propagated through the bottom layer of weights to the internal representation nodes. When a different stimulus input, such as a light, is presented (Fig. 2B) a different pattern of activation will emerge in the internal representation layer. Figure 2C shows that we can view geometric interpretation of an n-element internal representation as a point in n-dimensional *internal representation space* (analogous to a "psychological space" as described in Shepard, 1958, 1987).

Note that the location of points in this internal representation space is not uniquely determined by the stimulus input patterns; rather, these internal representations (points) are a transformation of the stimulus inputs by the bottom layer of weights in the network. Different associative weights result in a different placement of each stimulus pattern in the n-dimensional representation space. If these associative weights are modified during training, the location of stimulus patterns in the representation space will be altered.

Discrimination learning

Discrimination learning occurs through the additional development of associative weights in the top layer. These weights can be viewed as a mapping of points in the internal representation space into response categories. Figure 3A shows a sample classification task in which stimuli are represented as points in a two-dimensional internal representation space. Learning a set of weights to classify these patterns into two different "response" classes is analogous to finding a line that separates these two sets of input patterns. A random initial configuration of weights in the top layer is thus equivalent to a randomly placed line in this space. Training the top layer of weights incrementally moves this boundary line until it appropriately separates the stimulus input patterns (points in the space) into the appropriate response categories. If stimuli belonging to different classes are close together (as in Fig. 3A) the task is difficult because the line must be placed in precisely the right place.

If, however, the stimulus patterns are repositioned in the space (as in Fig. 3B, C) the task can be made easier for subse-

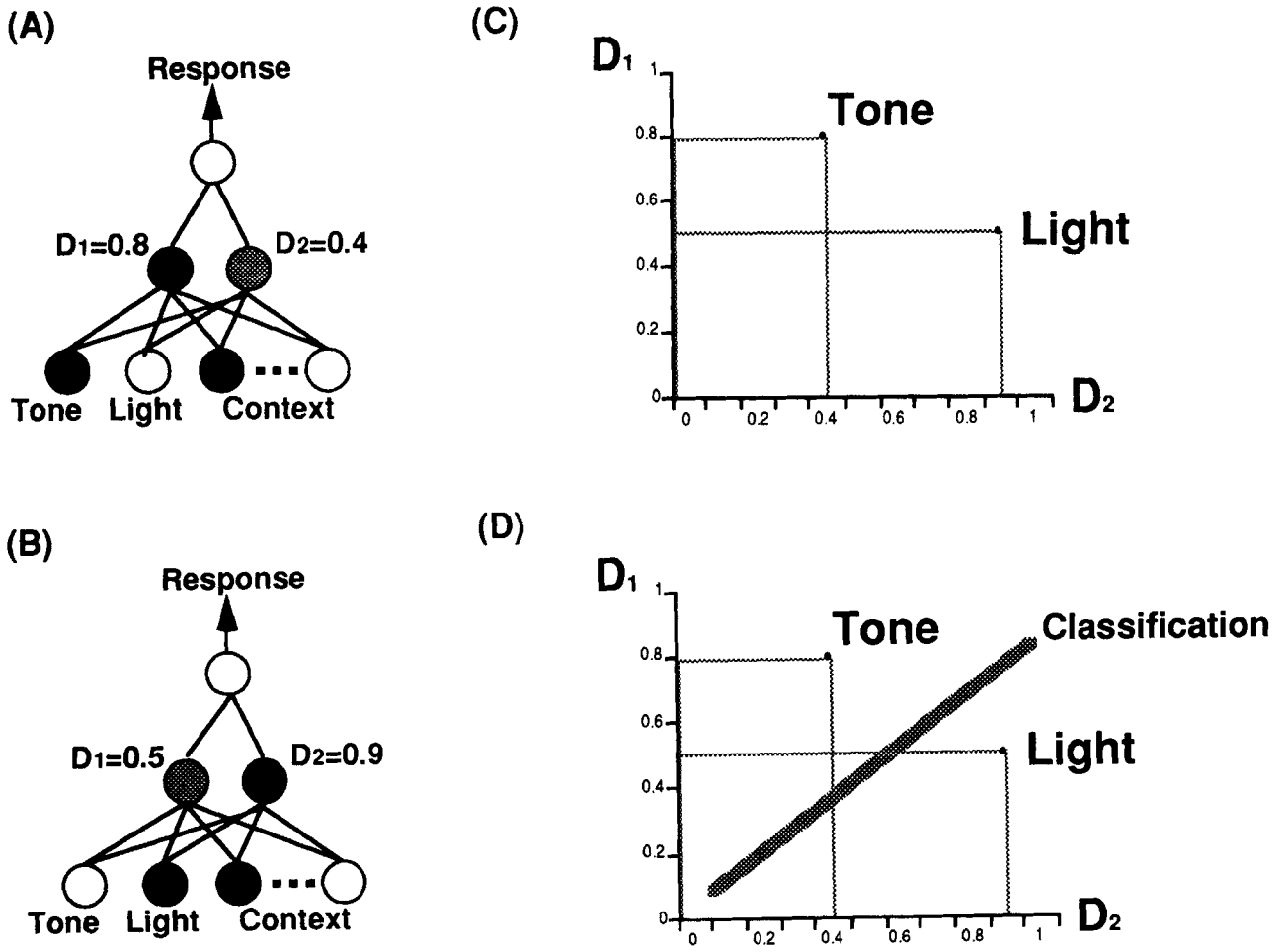


Fig. 2. Stimulus representation in a connectionist network with inputs representing stimulus cues (tone and light) and four contextual cues. (A) An example trial: tone (but not light) and some contextual cues are active in the input. Internal layer activations are a function of a weighted sum of these inputs; the internal representation formed in this example is (0.8, 0.4). The response, in turn, is a function of a weighted sum of these values. (B) Presenting a different stimulus input, light, results in a different internal representation: (0.5, 0.9). (C) These two-dimensional internal representations can be graphed in 2 dimensional internal representation space: each stimulus occupies a point corresponding to its current internal representation. The internal representations of tone and light from (A) and (B) are shown here. Changing the internal representation is equivalent to moving points within this space. (D) Learning to discriminate, or produce different responses to different stimuli, is analogous to partitioning internal representation space to separate the representations of stimuli into different response classes.

quent discriminations. With this example we see how an appropriate internal representation can facilitate a discrimination task by stretching the distance between stimuli that belong to different outcome categories. Note that some discrimination tasks may not be at all solvable with only a single classification boundary (e.g., if categories overlap) unless the patterns are repositioned in the representation space.

The transformation of representation space shown in Figure 3 is one example of a representational recoding that we believe depends on an intact hippocampal region. We propose that discrimination learning in the absence of an intact hippocampal region is analogous to using a fixed, nontransformed, internal representation (e.g., the initial representation space in Fig. 3A). As we show later in this paper, these hippocampal-dependent transformations of representation affect not only what can, and cannot, be learned, they also have strong implications for transfer effects to new tasks and for how animals will generalize from previous learning to novel stimuli.

The remainder of this article is organized as follows. First, our basic theory and model of hippocampal region function are described. We then show how this theory accounts for the performance of intact and hippocampal-lesioned animals in a wide range of testing conditions. Several novel predictions of the theory are described. We then relate this computational theory to prior interpretations of hippocampal function in elementary associative learning and argue that several prior qualitative characterizations of hippocampal function can be viewed as special cases of our theory. Finally, we discuss some limitations of the current approach and suggest directions for future empirical and theoretical research.

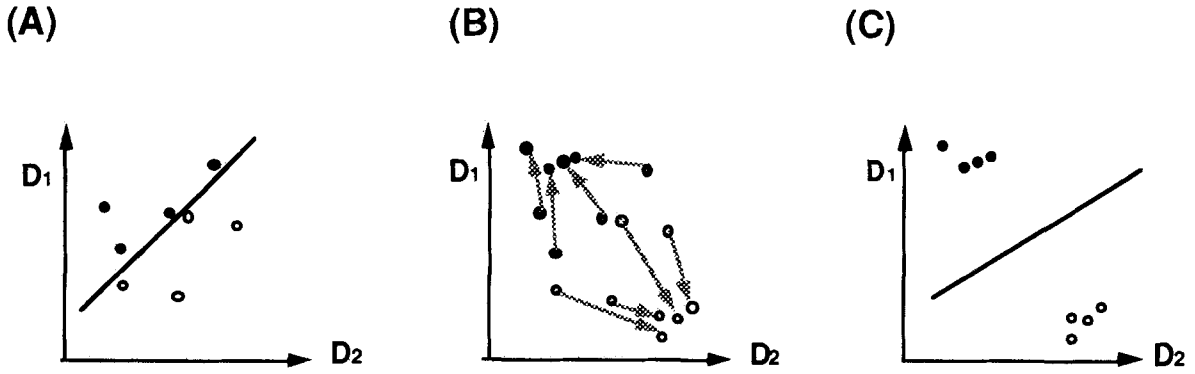


Fig. 3. Schematic view of movement in internal representation space. Stimuli are to be classified into two response categories (shown as black and white points). (A) Initial representations are very similar for stimuli in different classes; classification (represented by drawing a line that separates the representations) is difficult. (B) A new representation separates representations of stimuli that should be classified differently, while clustering members of the same class. (C) After this rerepresentation is formed, it is much easier to partition the classes successfully.

CORTICO-HIPPOCAMPAL MODEL

Central to our connectionist model is a network implementing the postulated hippocampal function of stimulus re-representation. An *autoassociator* network, as shown in Figure 4A, can be trained to reproduce its input pattern on its output nodes. When presented with a partial version of a previous stimulus, this network will try to reconstruct the complete input pattern on the output nodes. Thus, an autoassociator can function as a pattern-completion memory device that takes, as input, a noisy or partial stimulus pattern and outputs a previously stored input pattern (Anderson, 1977; Kohonen, 1984; McNaughton and Nadel, 1990). If the number of nodes in the internal layer of a multilayer autoassociator is smaller than the number of nodes in its input and output layers (as in Fig. 4A), the network is forced to generate a compressed internal representation (or encoding) of the stimuli on which it has been trained. This network is not only an autoassociator, but an *autoencoder* (Hinton, 1989).

If input signals on different input nodes (representing different input features) are uncorrelated across training patterns, autoencoding is not possible. If, however, there are correlated or redundant features, these can be compressed into a smaller, more compact, representation. The presence of correlated or redundant features indicates a nonrandom structure or statistical regularity across the training stimuli. For example, wings, feathers, water dwelling, and webbed feet are highly correlated as features of animals: e.g., water-dwelling winged animals tend to have feathers and webbed feet. Identifying this correlation and creating a compact representation of these features is analogous to a cluster analysis (or principal components analysis) that identifies the natural category of waterfowl.

An autoencoding network develops representations that are constrained by redundancy compression, one of the two representational biases our theory identifies with the hippocampal region. The second representational bias is predictive differentiation. We can augment the autoencoder by requiring that it produce output that not only recreates the input pattern, but also predicts the "outcome" (or classification) of the input pattern. An example of such a network, which we call a "predictive autoencoder," is shown in Figure 4B. The representations formed in the internal layer of a predictive autoencoder will still compress correlated or redundant features. However, they will also differentiate features that are especially predictive of the desired outcome as well. Thus, this network forms new internal representations that are biased by both predictive differentiation and redundancy compression.

Cortico-hippocampal interaction

Figure 5A shows the complete, intact cortico-hippocampal connectionist model. The hippocampal region network on the right is equivalent to the predictive autoencoder shown in Figure 4B. It receives sensory input and learns to recode stimulus information as described above. The network on the left receives the same stimulus input to its input layer; it is a two-layer network that learns to map from stimulus inputs to an appropriate response behavior. This network is presumed to be the site of long-term memory storage and to be responsible for response generation. As such, it is treated here as a very simplified abstraction of learning in cerebellar and cerebral cortices. A more complete version of the model might have several such cortical networks, all modulated by a hippocampal network (or networks). Only one cortical network is considered here.

While both networks in Figure 5A have two layers of weights, there is an important difference between the two networks. The hippocampal network on the right is presumed to have access to a learning procedure that allows recoding stimulus representations at the internal layer. An example of such a learning procedure is the backpropagation rule (e.g., Werbos, 1974; Parker, 1985; Rumelhart et al., 1986), used in the simulations to be described here. There are many other possible training procedures and the details of these procedures are not of great importance to our theory. The critical issue is how the network's architecture and task demands together constrain the allowable internal representations of stimulus patterns.

In contrast to the multilayer learning and stimulus recoding that occur in the hippocampal network, learning in the two

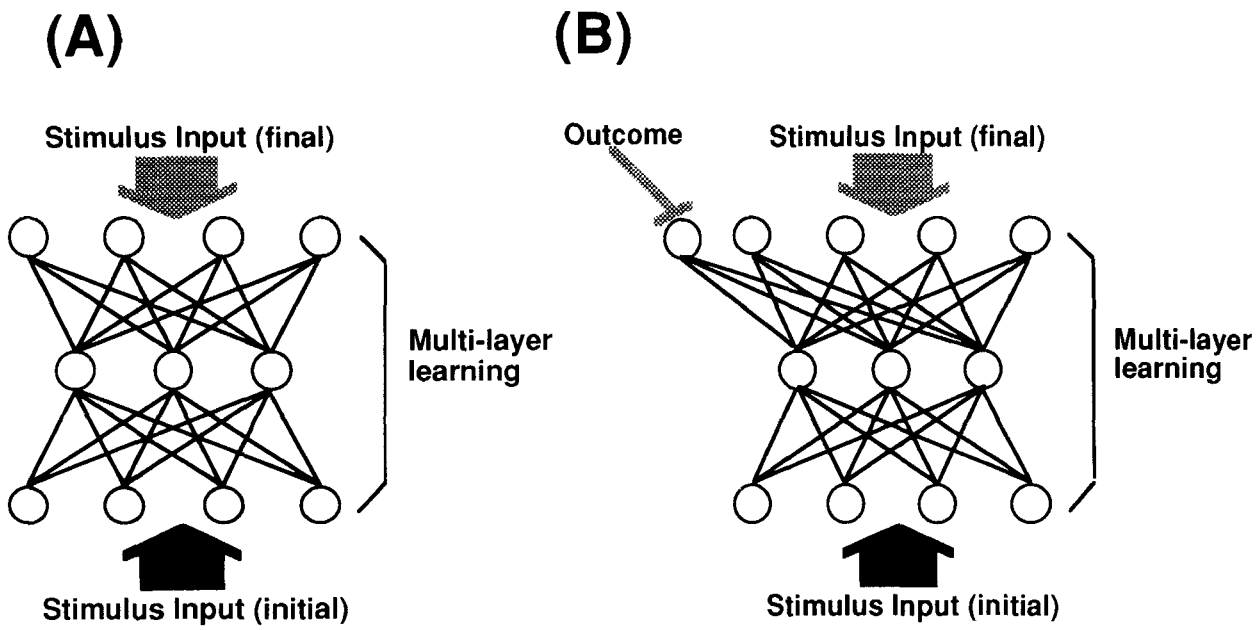


Fig. 4. (A) An autoencoder (Hinton, 1989). This network learns to reproduce its inputs at the output layer, using a multilayer learning algorithm. It contains a narrow internal layer, and so is forced to generate an internal representation that compresses redundancies in the input pattern. (B) A predictive autoencoder. This network is an autoencoder augmented with the constraint that it must also output a classification (or prediction of outcome) for the output pattern. The representation formed in the narrow internal layer must still compress redundancies; now, however, it must also differentiate representations of input features that are especially useful in predicting the outcome.

layers of the cortical network evolves independently. The bottom layer of weights (from input nodes to internal nodes) is trained so that sensory input signals generate internal representations similar to those being developed in the hippocampal network.[2] Independently and simultaneously, the top layer cortical weights learn to map from the evolving internal representation to an expectation of some relevant future outcome. These two independent learning processes in the cortical network are far simpler than the learning process used by the hippocampal network because they involve direct input-output mappings without discovering and constructing new stimulus representations. An example of this type of learning rule is the Rescorla-Wagner (1972) rule from animal learning theory, equivalent in this connectionist implementation to the LMS or "delta" rule (Widrow and Hoff, 1960). While more sophisticated than the correlational or Hebbian learning rules which have been used to describe synaptic plasticity (e.g., Stanton and Sejnowski, 1989; Brown et al., 1990), the Rescorla-Wagner/LMS rule can be built up from simple circuits using known synaptic modification rules (Donegan et al., 1989; Gluck and Granger, 1993).

The complete (intact) cortico-hippocampal model in Figure 5A can be shown to produce behavior comparable to normal animals. A lesioned version of this same model (Fig. 5B) can be compared with the behavior of hippocampal-lesioned animals. The lesioned model is produced by disabling the hippocampal network, thereby eliminating training information for the cortical network's internal representations. When deprived of hippocampal training signals, the cortical network can still adapt its upper layer of association weights but cannot modify its lower layer of weights to change its internal representation. It will still be able to learn classifications for which its current (now fixed) internal representation is sufficient.

In this article, we consider the domain of classical conditioning and show that the behaviors of the intact and lesioned systems are consistent with empirical data on behaviors in intact and hippocampal-lesioned animals. Before turning to these results, we briefly review classical conditioning and describe its instantiation within the model.

Modeling classical conditioning

Classical conditioning involves repeated pairing of a previously neutral stimulus such as a tone (the conditioned stimulus, or CS) with a response-evoking stimulus such as an airpuff to the eye (the unconditioned stimulus, or US). In most cases the US occurs after the CS onset. After sufficient paired presentations, the CS alone will evoke a conditioned response (CR), often similar to the unconditioned response (UR) evoked by the US. A conditioning paradigm that has received extensive analysis is the rabbit eyeblink response (see Gormezano et al., 1983) in which a tone or light CS is paired with an airpuff US, eliciting an eyeblink as both the UR and CR. The development and final form of the CR is influenced by

[2] Full details of the implementation are presented in the Appendix, but it is worth noting here that there need not be a one-to-one mapping from the hippocampal region internal nodes to the cortical network internal nodes. In fact, a sparse mapping suffices. As long as the cortical network internal layer learns to form a linear recombination of the hippocampal network rerepresentation, it is likely to have the same associative power as the hippocampal network.

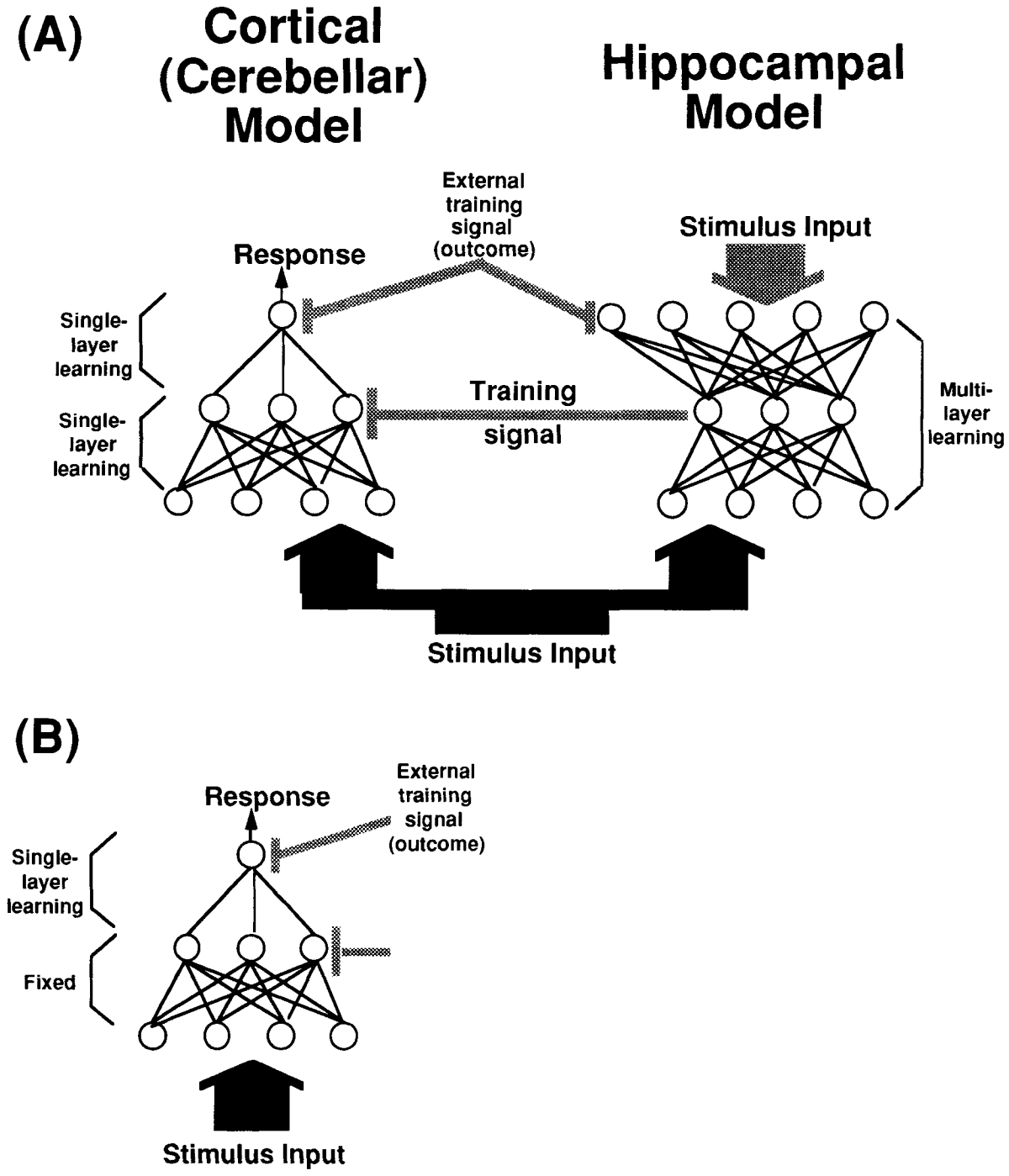


Fig. 5. The cortico-hippocampal model. (A) The hippocampal model (right) is a predictive autoencoder that learns to reproduce its stimulus inputs plus a prediction of the outcome. In the process it develops an internal representation constrained by predictive differentiation and redundancy compression. Cortical and cerebellar networks (one shown on left) also learn to map from stimulus inputs to a prediction of the output but they are not assumed to have access to a multilayer learning rule. The bottom layer of connections is trained so that stimulus inputs generate internal representations similar to those formed in the hippocampal model. Meanwhile, the top layer connections are trained to map from these evolving internal representations to an output that is interpreted as the behavioral response. (B) A lesioned model is constructed by disabling the hippocampal network, thereby eliminating training information to the cortical network lower layer of weights. These are now fixed, and the network is left with a static recoding in its internal layer. It can still learn by training its upper layer of connections to map from this fixed internal representation to a response.

parameters such as the strength and salience of the stimuli, the temporal interval between CS onset and US onset, and the reliability with which the CS predicts the US. We will concern ourselves here, however, with only the trial-level (e.g., nontemporal) properties of classical conditioning. Thus, we view the animal's task as the classification of sensory inputs according to whether or not they predict the US.

Stimulus inputs in our cortico-hippocampal model represent the state (present or absent) of all possible CSs and contextual cues. For motor-reflex conditioning, such as rabbit eyeblink conditioning, the site of long-term memory storage and response generation (the left network in Fig. 5A) would represent an abstraction of the learning that occurs in the cerebellar cortex, and possibly the underlying cerebellar nuclei (Thompson, 1986; Thompson and Gluck, 1990). This cerebellar cortical network model learns to produce a CR that predicts the occurrence of the US. In this way, CSs represent cues, while the CR is the system's response and the US represents the trial's outcome. We will generally use the terms *cue, response,* and *outcome* in preference to the more paradigm-specific CS, CR, and US to emphasize that classical conditioning is only one associative learning domain to which this general approach of cortico-hippocampal interaction might be applied.

An example of learning in this cortico-hippocampal model proceeds as follows (see Fig. 6). A pattern representing the input stimulus is presented to the input layers of both the cortical and hippocampal networks. This pattern represents the presence or absence of all stimulus cues, including both experimentally manipulated cues (e.g., tones and lights) and background contextual cues (e.g., the experimental cage, room, background noises, etc.). Contextual cues are always present but may change slowly over time, representing fluctuations in the conditioning environment and the animal's internal states. Both networks process the inputs simultaneously. Training occurs incrementally over many trials. The hippocampal network is trained to predict all sensory cues (including the US). The hippocampal network provides its current stimulus rerepresentation to the cortical network, which concurrently learns to reproduce this representation and associate this evolving internal representation with an expectation of the US—the trial outcome.

The example in Figure 6 shows a trial on which the tone input precedes US arrival. We will refer to this training situation using the notation ⟨Tone+⟩. Here, the animal's task is to learn to produce a response to the tone that predicts or anticipates the US. By contrast, we can compare this type of training task to another in which a cue (e.g., light) is presented but not followed by a US. The notation for this type of nonreinforced trial is: ⟨Light−⟩. Here, the animal should learn to withhold responding to the light cue. In a discrimination paradigm, the task is to learn to respond only during trials in which the cue (or cues) predict the US: for example, ⟨Tone+, Light−⟩. In each case, contextual stimuli are always present when the experimental cues are presented, and each presentation of conditioned cues is intermixed with several trials on

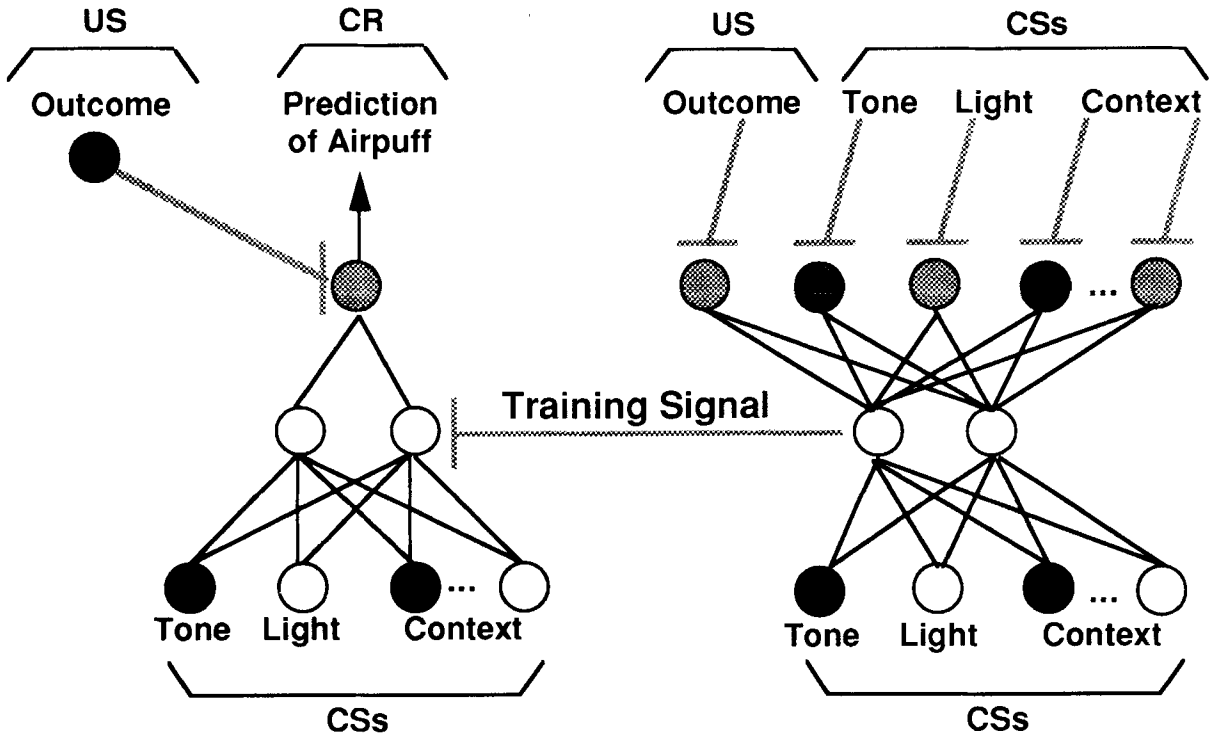


Fig. 6. An example ⟨Tone+⟩ conditioning trial: learning that the tone conditional stimulus (CS) predicts an airpuff unconditioned stimulus (US). Both networks receive the same stimulus inputs (tone CS on, light CS off, assorted contextual cues active). The hippocampal network learns to output these same values as well as a prediction of the outcome. The cortical network learns to reproduce the hippocampal network's current internal representation in its own internal layer, and then learns to map from this representation to a prediction of the outcome. This output is taken as the system's response.

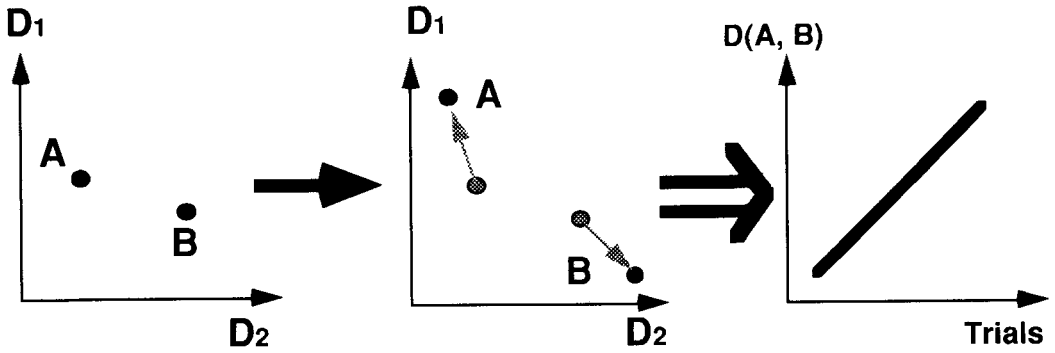


Fig. 7. (Left) Initial representations of two stimuli, A and B, with different outcomes, may undergo predictive differentiation (Center), which serves to move the representations further apart. (Right) Predictive differentiation may be monitored as an increase in the distance D(A, B) between the representations.

which the contextual cues alone are presented (with no US). These context-only trials will not be shown in the graphs reporting simulation results, but they are necessary to ensure that the system is learning selectively to respond to the conditioned stimuli, and not simply to respond on every trial.

We now turn to evaluating our theory and model of hippocampal region function in classical conditioning. First, we consider the implications of predictive differentiation, comparing the behavior of the intact and lesioned model to available empirical data on a variety of training procedures sensitive to this representational bias. We then do the same for redundancy compression, showing how our model accounts for a wide range of hippocampal-dependent behaviors and predicts several novel (and still untested) effects of hippocampal lesions.

EMPIRICAL TESTS: I. PREDICTIVE DIFFERENTIATION

As described earlier, predictive differentiation increases the discriminability (within the internal representation) of stimuli with different predictive consequences. By allocating more resources to relevant stimulus dimensions, the representation of these dimensions becomes, in effect, stretched. Stimuli that differ along these dimensions will accordingly move further apart in representation space (see Nosofsky, 1984, for a related analysis of human classification learning). This stretching of stimulus space along relevant dimensions was previously illustrated in Figure 3.

In analyzing our model's performance—and the expected role of predictive differentiation—it is technically possible to keep track of individual stimulus representations as they change during learning. However, these specific patterns of activity along the internal units are high-dimensional, distributed, and not easily interpreted. What is more informative, however, are the relationships among these stimulus representations and how these relationships change during learning. In particular, it is informative to track the changes in distance between the representations of relevant training stimuli. For example, two stimulus patterns A and B might initially lie in representation space as shown in Figure 7, Left. If A and B have different predictive consequences, we expect A and B to move further apart in representation space, as illustrated by Figure 7, Center. Rather than plotting this movement directly, we can measure the distance between A and B in the internal representation space, D(A, B), and graph changes in this distance during training trials (Fig. 7, Right). These changes reflect alterations in the internal representation space. Because the hippocampal-lesioned model is presumed to have fixed internal representations, D(A, B) will change for the intact model only. In the analyses to follow we will often plot the changes in D(A, B) that occur during training on specific conditioning tasks; this will better allow us to interpret how and why the model behaves as it does.

What are the behavioral consequences of this altered stimulus representation? Predictive differentiation will be most readily apparent—at the behavioral level—in training situations in which the altered representation space either facilitates, or retards, subsequent transfer to a new discrimination task. After describing the effects of predictive differentiation on ⟨A+, B−⟩ discrimination learning, we will examine several such transfer tasks, including reversal learning, easy-hard transfer, and stimulus generalization.

Stimulus discrimination

One of the most elementary learning tasks is stimulus discrimination: learning, for example, that CS A predicts the US while CS B does not (i.e., ⟨A+, B−⟩ training). Figure 8A shows the development of an appropriate CR to A and the absence of a CR to B for both the intact and lesioned network models.[3] The lesioned system learns this discrimination through the adaptive modification of the cortical network's upper level of weights. The cortical network's internal representation remains fixed for the lesioned model because the lower weights are not modified in the absence of training signals from a hippocampal network. In contrast, the intact model learns to enhance the discriminability of the two stimuli by pulling them apart in internal representation space (Fig. 7). This representational recoding is reflected by the evolution of D(A, B) in the intact model, where it increases as training progresses (Fig. 8B).

Turning back to Figure 8A, we observe that the lesioned network learns the discrimination faster than the intact net-

[3] Each data point in the figures represents the conditioned response, averaged over 10 simulations, after the given number of learning blocks. Distance $D(x, y)$ between the representations for stimuli x and y is calculated from the cortical network internal layer representations using a city block metric.

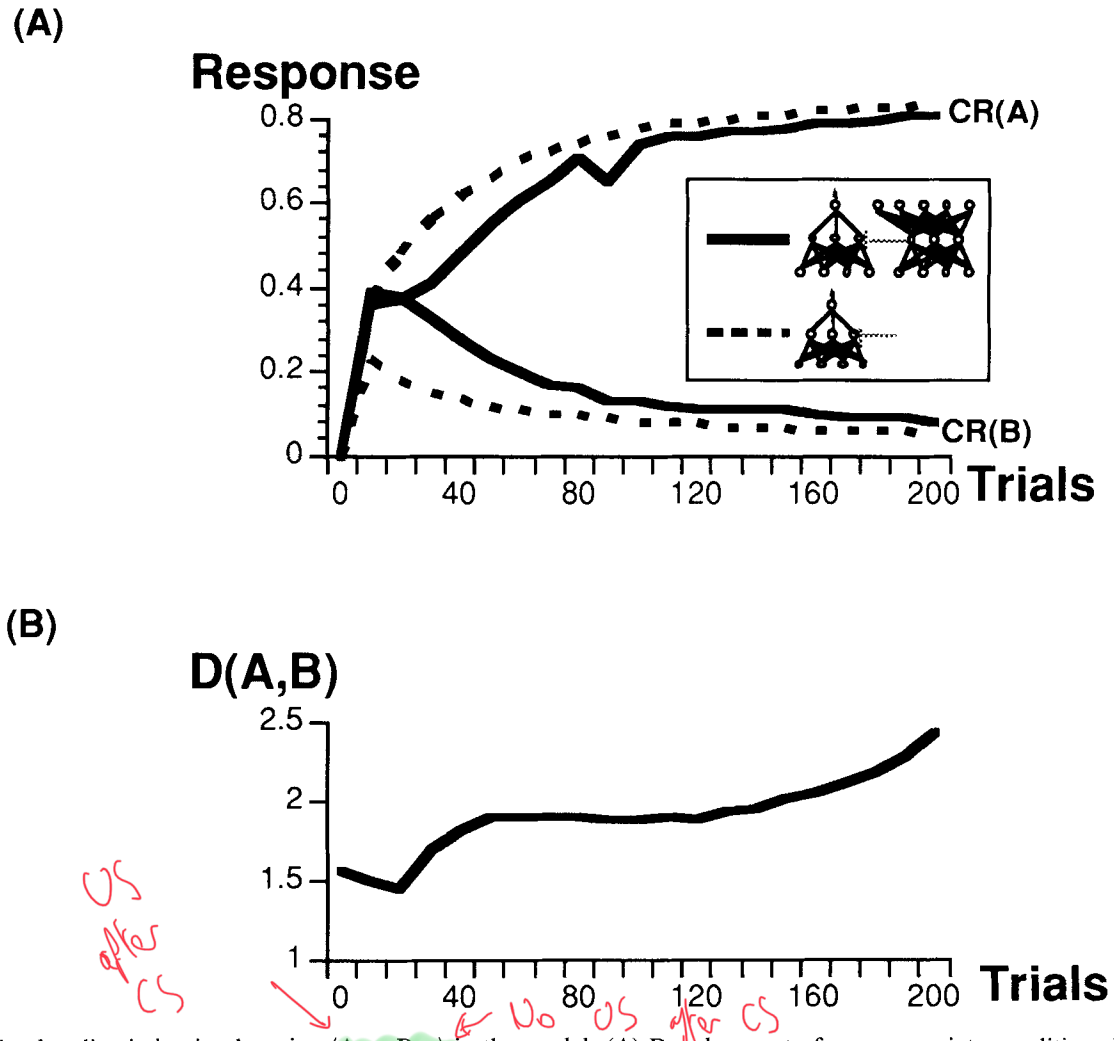


Fig. 8. Stimulus discrimination learning ⟨A+, B−⟩ in the model. (A) Development of an appropriate conditioned response (CR) to A but not to B in the intact (solid line) and lesioned (dashed line) models. The lesioned system, in which the cerebellar network can make use of its fixed internal representation, learns slightly more quickly. The intact system, in which the cerebellar network must acquire the hippocampal network's internal representation, learns more slowly. (B) The progressive increase in D(A, B) during training in the intact model reflects predictive differentiation in the stimulus representations. Simulation results in (A) are consistent with the empirical findings of Port et al. (1985), Eichenbaum et al. (1988), Schmaltz and Theios (1972), and others.

work. Why does our model expect this? Note that the lesioned network adjusts its top layer of classification weights based on a static (constant) internal representation. In contrast, the intact system attempts to learn a classification based on a changing representation. Thus, only when the stimulus recoding is completely acquired from the hippocampal model can the final classification weights be learned. This, naturally, results in slower learning for the intact model.

Careful comparisons of intact and lesioned animals have often shown this same effect: a facilitation of the rate of stimulus discrimination learning for hippocampal-lesioned animals (Schmaltz and Theios, 1972; Port et al., 1985; Eichenbaum et al., 1986, 1988, 1991).[4] Some researchers have interpreted these experimental results as suggesting that the hippocampus impairs learning of stimulus discriminations. For example, one early paper noted the selective facilitation of CR acquisition in hippocampal-lesioned animals and concluded that, under optimal conditions, "the modulatory process [of the hippocampus] may delay the initial occurrence of CRs. At an optimal interstimulus interval, modulation may be minimal or unnecessary. Hence, hippocampectomy has no significant effect" (Port et al., 1986).

Our theory leads to a different interpretation of these empirical data. The learned alterations in representation implied by the observed changes in D(A, B) in Figure 8B suggest that slower learning for intact animals is a result of these animals learning *more* than what is learned by the lesioned animals. In particular, the stimulus recoding learned by the intact model can be interpreted as learning what cues are *relevant*, in addition to learning the predictive consequences of these

[4] Most empirical studies referenced for comparison with the model involve rabbit NMR (eyeblink) conditioning; however, where such studies were unavailable, work in rat or pigeon is cited.

cues. In contrast, the lesioned model learns only about the predictive consequences of all the cues. This distinction between learning what cues are relevant and learning what those cues predict has a long history in psychology and allows us to interpret several experimental studies involving transfer of learning from one task to another. In each case, the modified stimulus representations resulting from the first learning task have predictable consequences for the system's behavior in the transfer task. Hippocampal-lesioned animals, however, are not expected to show these same transfer effects.

Discrimination reversal

Our theory suggests that the internal representations of predictive stimuli are pulled apart during discrimination training. If subsequent tasks can make use of this altered representation, learning will be facilitated. The simplest example of this expected facilitation is reversal learning. A reversal task involves an initial phase of learning ⟨A+, B−⟩ followed by a phase of learning ⟨A−, B+⟩.

With an intact hippocampal region, learning ⟨A+, B−⟩ results in pulling apart stimulus representations for A and B—thus D(A, B) increases. When the task switches to ⟨A−, B+⟩, the stimulus representations are already very distinct—only the new discrimination must be learned. In contrast, our theory expects that without an intact hippocampal region, reversals should not be facilitated because no stimulus rerepresentations have been formed. Furthermore, there should be an impairment in learning the reversed task because the old classification knowledge must be "unlearned" before the new task can be acquired.

Figure 9 shows simulation results from the model that illustrate these implications of the theory. The intact model takes slightly fewer trials to learn the reversed task. In contrast, the lesioned model has great difficulty with all of the reversals—taking more time to learn the reversal than to acquire the initial discrimination. Figure 9 also shows that the lesioned model learns the first task faster than the intact model—the same facilitation of stimulus discrimination shown in Figure 8A.

The consequences of predictive differentiation for reversal learning seen in the simulation results of Figure 9 have been observed in experimental studies. Intact animals show a progressive improvement on successive reversals in a wide range of training procedures and paradigms (see Sutherland and Mackintosh, 1971, for review). In contrast, hippocampal-lesioned animals show profound impairment in learning to reverse a trained discrimination (Berger and Orr, 1983).

Easy-hard transfer

Additional evidence for the representational stretching caused by predictive differentiation can be found in studies of learning about stimulus values that vary along a physical (or psychophysical) continuum such as tone frequency. For example, a task might be to differentiate two stimuli with widely differing values along a stimulus continuum (e.g., a very high and a very low tone). Our theory expects that learning this task should stretch the internal representation of that continuum. This stretching should then facilitate subsequent discrimination of other stimuli along the same continuum (e.g., moderately low and moderately high tones).

For example, an easy task for the model is to learn to respond when stimulus A has a value of 0.9, but not when it has a value of 0.1: ⟨A = 0.9+, A = 0.1−⟩. Following training, the distance between the representations activated by A = 0.9 and A = 0.1 should increase, stretching apart all feature values along this stimulus continuum (see Fig. 10A). Figure 10B shows simulation results from our model illustrating how prior training on an easy discrimination ⟨A = 0.9+, A = 0.1−⟩ facilitates learning a second, harder discrimination of intermediate stimulus values (e.g., ⟨A = 0.6+, A = 0.4−⟩). Learning the harder second task is much quicker for this pretrained group than for a control group that was not pretrained on the easy ⟨A = 0.9+, A = 0.1−⟩ discrimination. In fact, training on the easy task facilitates learning the hard task more than an equal amount of training on the hard task itself (simulations not shown).

This phenomenon is well known in the experimental literature on animal learning and is referred to as easy-hard learning, or transfer along a continuum (Lawrence, 1952; Riley, 1968). It is also related to the "errorless" learning paradigms described by Terrace (1963). In the intact animal, prior training on an easy discrimination (extreme values) facilitates learning of a hard discrimination (intermediate values) more than equivalent amounts of training on the hard discrimination alone (Lawrence, 1952).

Our cortico-hippocampal theory makes a novel prediction about easy-hard learning after hippocampal lesion. Figure 10C shows the performance of the lesioned model: after learning the easy task, there is only a small amount of facilitation of learning the hard task, due to stimulus generalization. We predict there will likewise be no easy-hard facilitation in hippocampal-lesioned animals relative to control levels of stimulus generalization.

As noted above, one factor in easy-hard facilitation is stimulus generalization. After learning ⟨A = 0.9+, A = 0.1−⟩, simple generalization of learned responses to nearby stimuli might account for some of the facilitation in learning ⟨A = 0.6+, A = 0.4−⟩. However, our theory also expects a strong facilitation in the intact model when the contingencies are reversed in the second phase of training. For example, the easy task might be to learn ⟨A = 0.9+, A = 0.1−⟩; the hard reversal would then be to learn ⟨A = 0.4+, A = 0.6−⟩. Stimulus generalization would expect that learning the easy task should impair the hard reversal. Our model predicts that, like in the standard easy-hard shift, dimension stretching in the easy task makes the cues more discriminable in the second, hard reversal task. The intact cortico-hippocampal model does show this facilitation in the hard reversal (simulations not shown). Consistent with our theory's expectations, Mackintosh and Little (1970) found that intact animals showed the facilitation for the easy-hard reversal shift.

These analyses of the easy-hard reversal shift lead to a more rigorous second novel prediction of our theory. Hippocampal lesions should eliminate facilitation if the hard task is a reversal of the easy task—learning the easy task may even impair learning the hard reversal. Not only does the lesioned model lose any facilitation gained from training in the initial easy task, but stimulus generalization should cause a strong impairment in learning the hard, reversed task (simulations not shown). We predict the same loss of facilitation in hippocam-

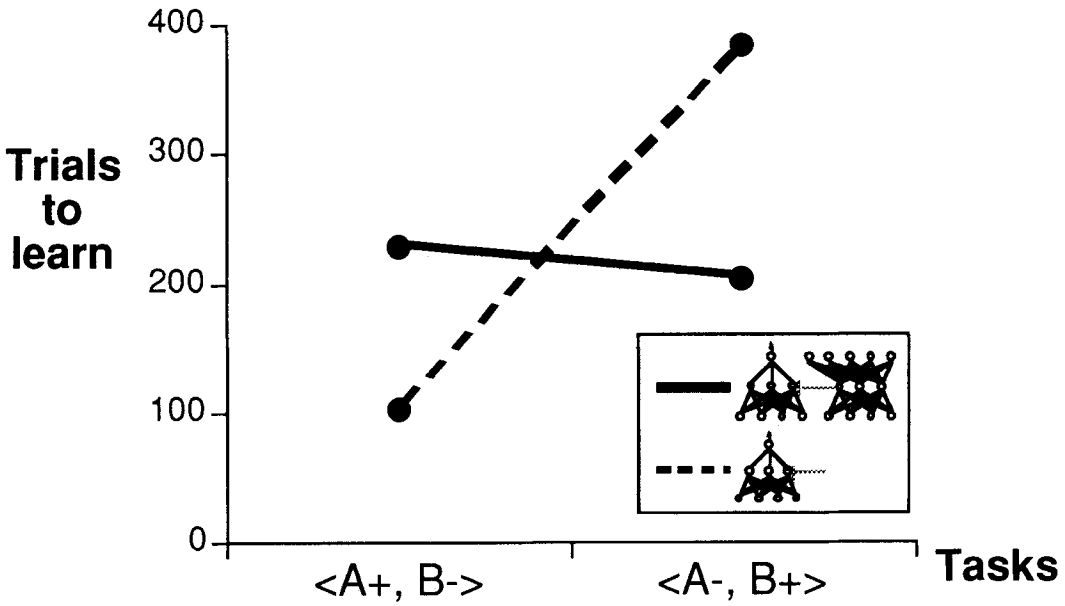


Fig. 9. Reversal learning in the intact (solid line) and lesioned (dashed line) models. In the first learning phase, the lesioned system learns a stimulus discrimination task ⟨A+, B−⟩ faster than the intact system. In a second phase the task valences are reversed to ⟨A−, B+⟩. The intact system already has differentiated representations for A and B from the first phase and must learn only the new classifications. Learning is facilitated. In the lesioned system, however, no stimulus rerepresentations are formed during the first phase. Learning in the second phase is retarded because old classification knowledge must be completely unlearned before the new task can be acquired. The retardation of reversal learning after hippocampal lesion is consistent with the empirical findings of Berger and Orr (1983).

pal lesioned animals. As noted earlier, impairment of reversal learning is characteristic of hippocampal-lesioned animals (Berger and Orr, 1983).

Stimulus generalization

Stimulus generalization studies provide another source of data on the representation of stimulus continua. In a common stimulus generalization experiment, a subject receives training trials with a single stimulus and is subsequently tested with novel stimuli that vary along a stimulus continuum from the training stimulus. For example, an animal might first be trained to associate an 800 Hz tone with a subsequent shock US. The animal would then be tested with novel similar cues that vary along this continuum, for example, tones of 700 Hz, 900 Hz, 1,000 Hz, etc.

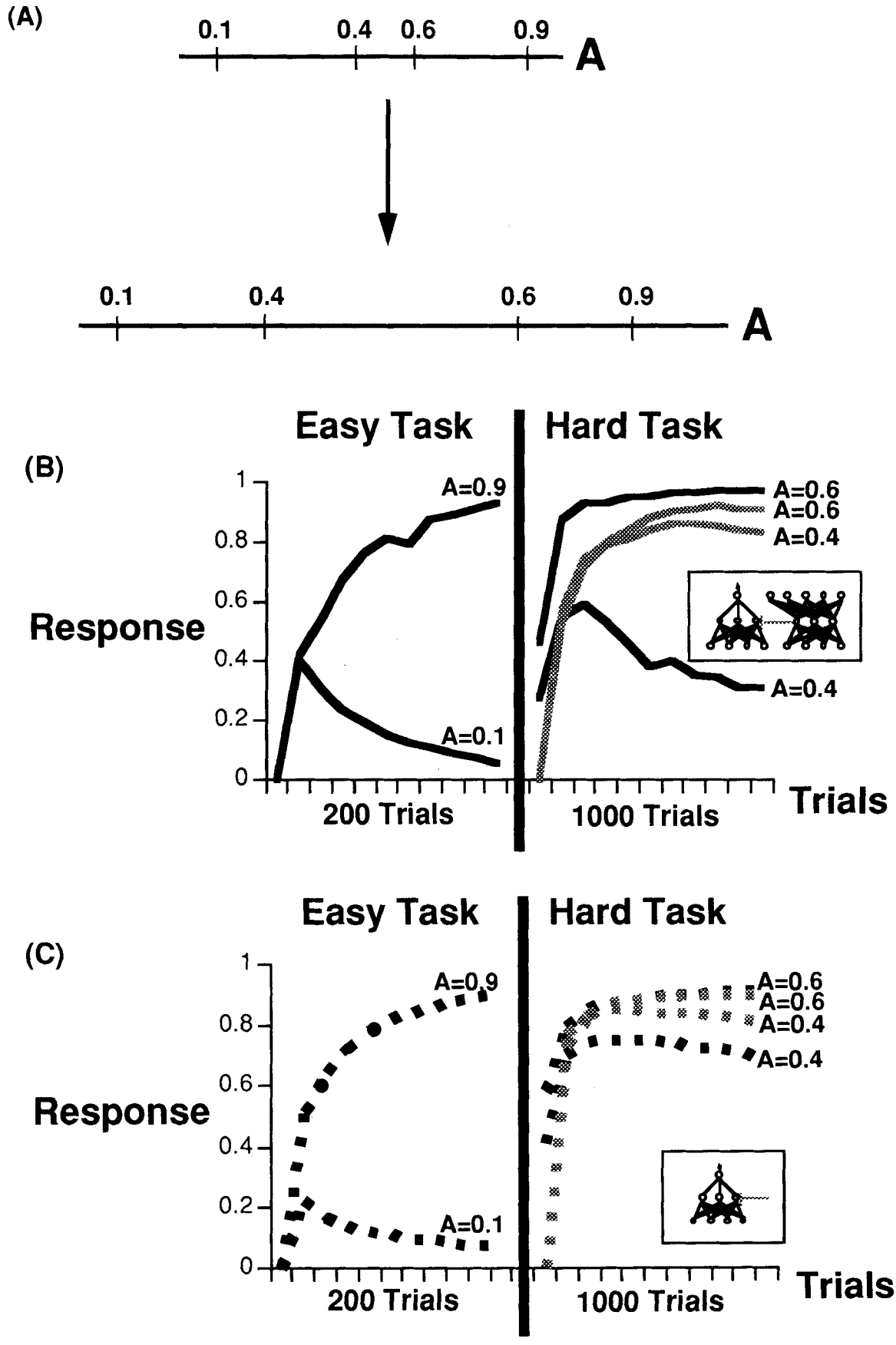
A well-established principle of stimulus generalization across a wide range of training procedures is the relationship between the probability (or strength) of generalization and psychological distance: the curve is concave-upwards, approximating an exponential-decay curve (Shepard, 1958, 1987). This principle has received additional support within mathematical models of human pattern classification and categorization (Nosofsky, 1984; Gluck, 1991).

We see appropriately shaped stimulus generalization gradients in simulations with our intact model (Fig. 11A). The steep gradient near the training stimulus occurs because the effect of predictive differentiation is to enhance the difference between the training stimulus and all other cues. A consequence of pulling the training stimulus further away from all other stimuli is to decrease the impact of stimulus generalization between the training stimulus and other stimuli.

Thus, our theory expects that in the absence of this hippocampal-dependent alteration of representation space, there should be more stimulus generalization in the lesioned model than in the intact model (Fig. 11A). The empirical studies of Solomon and Moore (1975) showed this same effect: hippocampal-lesioned animals exhibit broader stimulus generalization gradients than intact control animals.

We can also look at the development of the stimulus generalization gradient during ⟨A+⟩ training. Figure 11B shows

Fig. 10. Easy-hard transfer, or transfer along a continuum. (A) Initially, discriminating A = 0.1 and A = 0.9 is rapid (easy), since the stimuli are distinct; in contrast, discriminating A = 0.4 from A = 0.6 is much slower (harder). During learning of the easy task, however, predictive differentiation will result in "stretching" of representation space: the representations of A = 0.4 and A = 0.6 will also be pulled apart. The result is that the hard task will be facilitated after pretraining to the easy task. (B) The intact model shows this effect. Learning the hard discrimination ⟨A = 0.6+, A = 0.4−⟩ after prior training on the easy discrimination ⟨A = 0.9+, A = 0.1−⟩ (black line) is faster relative to learning the hard task alone (gray line). (C) In the lesioned model, with no stimulus rerepresentation and hence no predictive differentiation, the effect disappears. Learning the hard task alone (gray line) is only slightly slower than if pretraining to the easy task occurs (black line). The effect disappears completely if the valences on the hard task are reversed to ⟨A = 0.1+, A = 0.9−⟩. Simulation results in (B) are consistent with the empirical findings of Lawrence (1952); results in (C) represent a novel prediction of the model.


(A)
0.1
0.4
0.6
0.9
A
0.1
0.4
0.6
0.9
A
(B)
Easy Task
Hard Task
Response
1
0.8
0.6
0.4
0.2
0
A=0.9
A=0.1
A=0.6
A=0.6
A=0.4
A=0.4
200 Trials
1000 Trials
Trials
(C)
Easy Task
Hard Task
Response
1
0.8
0.6
0.4
0.2
0
A=0.9
A=0.1
A=0.6
A=0.6
A=0.4
A=0.4
200 Trials
1000 Trials
Trials

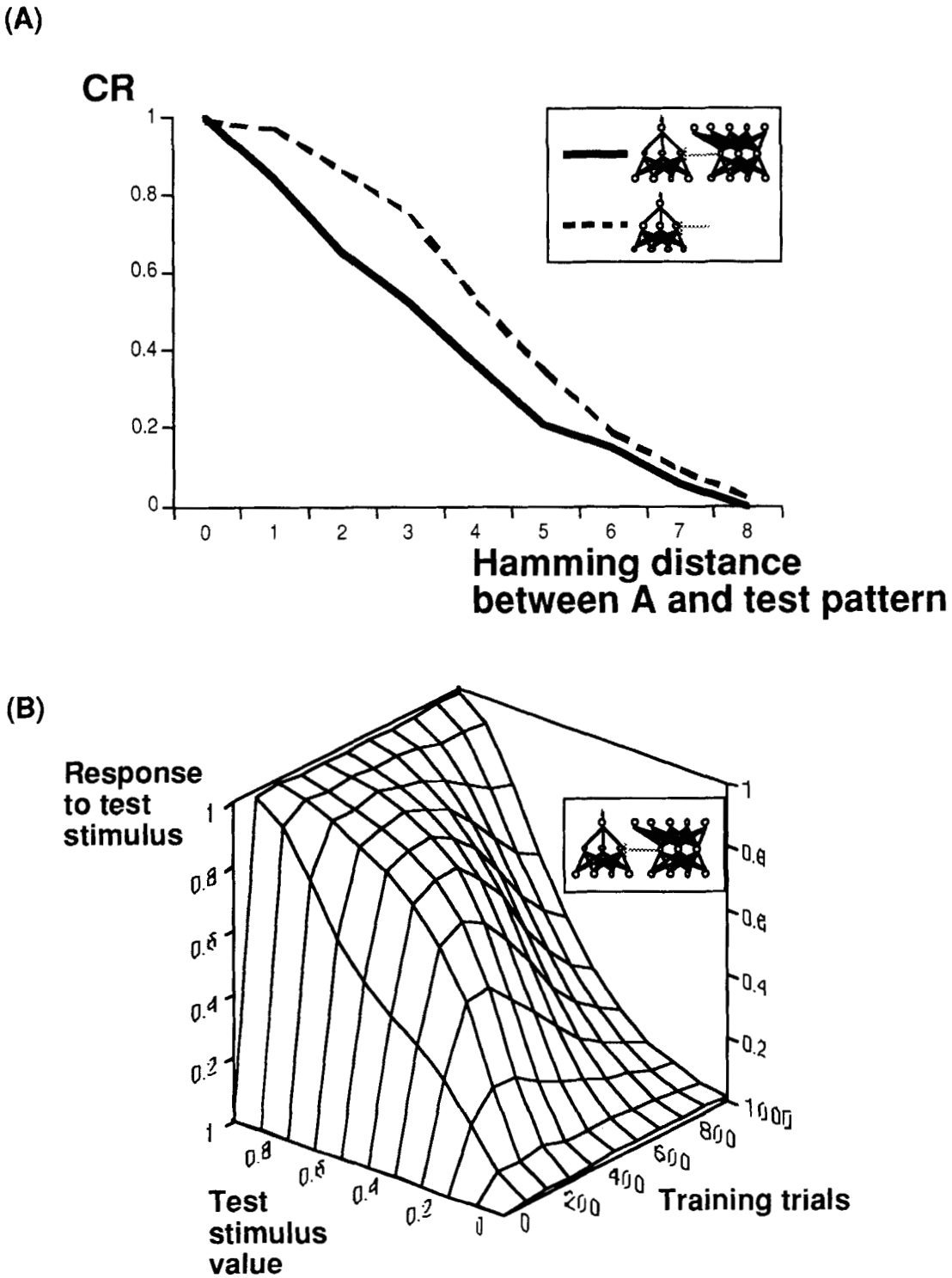


Fig. 11. Stimulus generalization gradients. (A) A stimulus generalization gradient is constructed by training the system to one stimulus input, such as ⟨A+⟩, and then testing the response to various similar inputs, e.g., input patterns at various hamming distances from A. A hamming distance counts the number of input units that are different in the test pattern compared to the training pattern. The intact model (solid line) shows an approximately exponential decay of response with distance to the training input. The lesioned model (dashed line) shows a much broader generalization gradient: it is more likely to produce a response to stimuli near A in representation space compared to the intact model. Simulation results are consistent with the empirical findings of Solomon and Moore (1975). (B) Development of the stimulus generalization gradient during ⟨A+⟩ training in the intact system is nonmonotonic. An initially roughly linear gradient develops into a broadly tuned curve in the early stages of learning. This is gradually fine-tuned to the very steep gradient shown in (A). Simulation results (shown for a single experimental run) are consistent with empirical findings of Thompson (1958).

that, in the initial stages of training, the generalization gradient in an intact system is roughly linear; this quickly develops into a broadly tuned curve. It would be expected that, at this stage of training, the system would show high generalization to patterns at various distances from A. With more training, this gradient is progressively fine-tuned, until the gradient is very steep. After overtraining, therefore, the system will show strong response only to test patterns that are very similar to A. This result is consistent with studies by Thompson (1958), who found a nonmonotonic development of the generalization gradient: an initial phase of overgeneralization was succeeded by one of very narrow generalization. Analogous simulations with the lesioned model (not shown) indicate a more nearly monotonic development of the stimulus generalization gradient. We know of no data on the development of generalization gradients in hippocampal-lesioned animals. Precise predictions about the expected form and time course of these gradients will depend, however, on specific training parameters.

EMPIRICAL TESTS: II. REDUNDANCY COMPRESSION

As described in the previous section, hippocampal-dependent enhancement of predictive differentiation results in stimulus representations moving apart if stimuli have different predictive values. Redundancy compression has the opposite effect: moving stimulus representations closer. As discussed earlier, redundancy compression is a bias to efficiently and compactly encode stimuli using minimal resources while still retaining the ability to differentiate between distinct stimulus patterns. It has the effect of combining or clustering correlated (and hence redundant) stimulus features. As a result, the representations of stimuli that differ on these redundant features will move closer together. Figure 12 illustrates the effect of redundancy compression on the representation of two stimuli with redundant cues.

As with predictive differentiation, the clearest behavioral implications of redundancy compression are for transfer tasks. In this section we consider several training procedures in which stimulus representations are altered by redundancy compression in an initial training task. Learning and generalization performance on subsequent transfer tasks is either facilitated or retarded, depending on whether or not the initial redundancy is maintained or altered.

Latent inhibition

Our theory expects that unreinforced stimulus preexposure will lead to compression of the representation of that stimulus with background contextual cues. For example, during a series of unreinforced presentations of stimulus A ($\langle A- \rangle$ preexposure), the presence of A is irrelevant to predicting a US. With extended $\langle A- \rangle$ preexposure, the distance between the representations activated when A is present and when A is absent (context alone) will move closer together, as shown in Figure 13B. If a subsequent training procedure requires the learner to distinguish between the cue and the context ($\langle A+ \rangle$ training), we expect the initial $\langle A- \rangle$ preexposure (and stimulus recoding) to impair learning the new task because the system must now respond to a feature it previously learned to ignore by compressing it with the contextual stimuli. Figure 13A shows that $\langle A+ \rangle$ learning by the intact model is slowed after unreinforced preexposure to $\langle A- \rangle$ compared to a control group that was not preexposed to A. As expected by our theory, this phenomenon, called latent inhibition, is seen empirically in a wide range of conditioning procedures (e.g., Lubow, 1973).

Because latent inhibition in the intact model depends on hippocampal-dependent stimulus recoding, it is not expected in the lesioned model. No matter how irrelevant A may be in the preexposure phase, there is no mechanism in the lesioned model for constructing a representation that minimizes the salience of A. We expect, therefore, that learning the $\langle A+ \rangle$ task should be just as fast in the preexposed lesioned condition as for lesioned control animals (no preexposure). Figure 14 shows an absence of latent inhibition in the lesioned model. Consistent with these simulation results, hippocampal lesions have been shown to eliminate latent inhibition in rabbit eyeblink conditioning (Solomon and Moore, 1975).

Implicit in our interpretation of latent inhibition is the assumption that context plays a key role. Latent inhibition, in our theory, is a result of the representation of A being compressed with the representation of the background context.

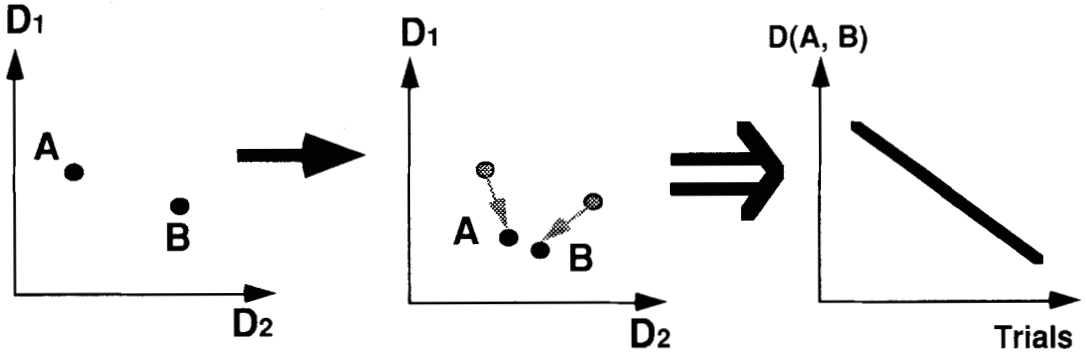


Fig. 12. (Left) Initial representations of two stimuli, A and B, with similar outcomes, may undergo redundancy compression (Center), which serves to move the representations closer together. (Right) Redundancy compression may be monitored as a decrease in the distance D(A, B) between the representations.

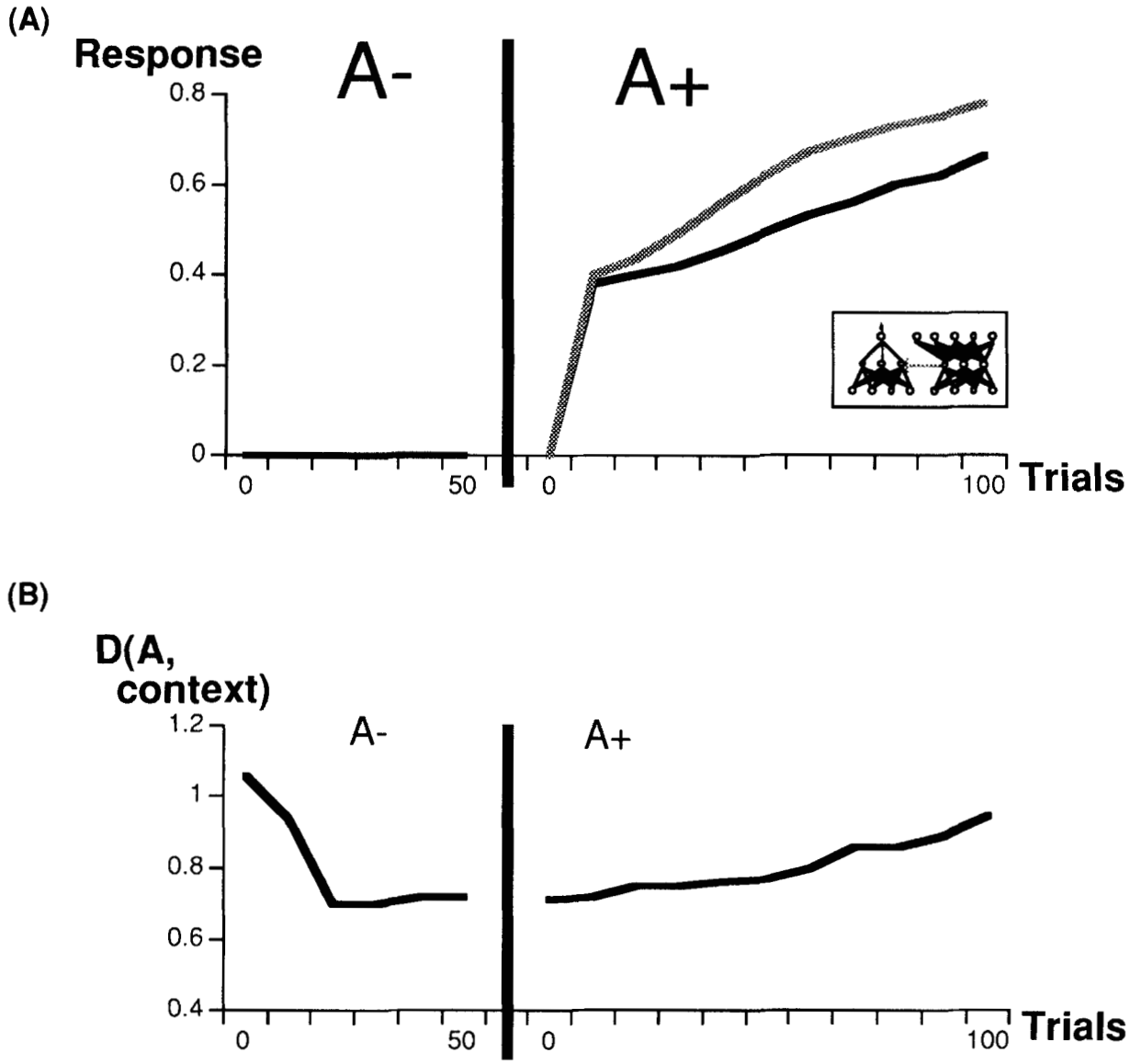


Fig. 13. Latent inhibition in the intact model. (A) Preexposure to ⟨A −⟩ retards ⟨A +⟩ learning (black line), relative to learning ⟨A +⟩ with no preexposure (gray line). (B) During ⟨A −⟩ preexposure neither A nor the constant contextual cues are predictive of US arrival. Redundancy compression decreases D(A, context), since A is no more salient a predictor than the contextual cues. During ⟨A +⟩ training, A becomes a good predictor of US arrival. Predictive differentiation occurs, and D(A, context) increases—undoing the redundancy compression of the first phase. This slows learning relative to the condition in which no preexposure occurs. Simulation results in (A) are consistent with empirical findings of Lubow (1973).

If the context is switched between the end of ⟨A −⟩ preexposure and the start of ⟨A +⟩ training, we would expect to see a release from latent inhibition. Because the representation of A has not been compressed with this new context, we expect no impairment in learning to discriminate A from this new context. Simulations (not shown) confirm this expectation: there is no effect of ⟨A −⟩ preexposure with the intact model on ⟨A +⟩ training, given a context shift between learning phases. Consistent with the model, experimental studies have shown that when intact animals are shifted from one context to another between the ⟨A −⟩ preexposure and ⟨A +⟩ training phases, there is no latent inhibition (Reiss and Wagner, 1972).

Sensory preconditioning

Whenever two cues co-occur to the extent that the presence of one is a good indicator that the other will be present, they can be combined or compressed within the stimulus representation. For example, in compressing (standard) English text it is unnecessary to encode both components of the sequence "qu"; representing the "q" is enough to infer that a "u" has also appeared. It is precisely this type of statistical redundancy that allows for efficient data compression in computer memories and in high-speed computer modems that rapidly transmit data across telecommunication lines.

Redundancy compression allocates representational resources efficiently. As suggested above, it can also affect

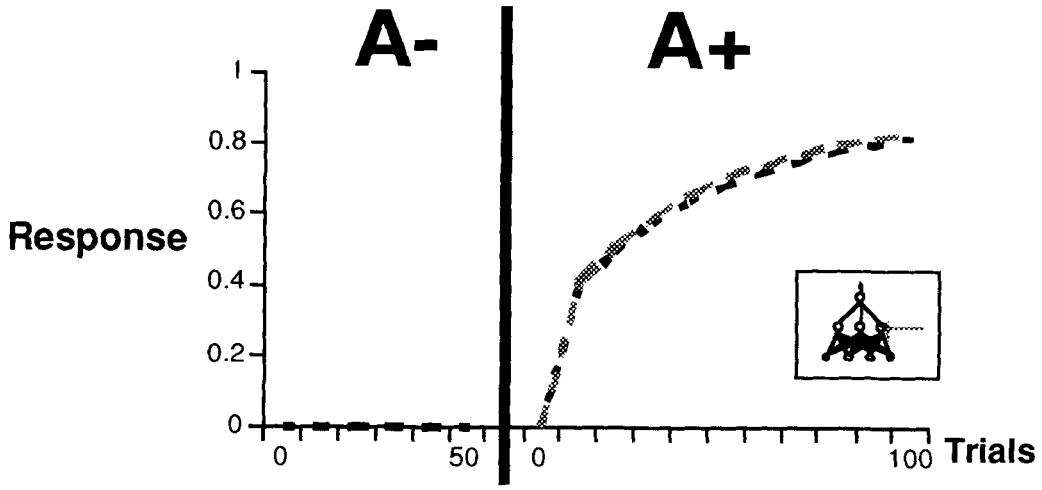


Fig. 14. Absence of latent inhibition in the lesioned model. There is no redundancy compression during preexposure to ⟨A −⟩ (black line), and therefore ⟨A +⟩ learning is not slowed relative to the control condition of no ⟨A −⟩ preexposure. Simulation results are consistent with the empirical findings of Solomon and Moore (1975).

transfer from one learning task to another. If two cues are compressed within a representation then they will be treated as a unit; learning about one will transfer to the other to the extent that their representations are compressed. Using the example from a previous section, this would be analogous to applying knowledge about birds with webbed feet to birds that live in the water. This transfer can be interpreted as assuming that both features are diagnostic of a common class of animals (i.e., waterfowl).

As in our previous analyses of latent inhibition, the behavioral consequences of this stimulus recoding can most clearly be seen by altering the correlational status of stimulus features from one training situation to another. For example, if an animal is first given unreinforced trials with stimuli A and B presented together as a compound cue (⟨AB −⟩), our theory expects the representations of the individual cues to be compressed in an intact animal. If the animal is then trained that A (alone) predicts the US, some of this association will transfer to B. Thus, the animal should respond to test presentations of B more strongly than if the initial compound preexposure training had not occurred (see Fig. 15). Again, our theory expects that this effect will not be present in hippocampal-lesioned animals because there is no new representation formed during the preexposure phase.

Empirical studies of this same training procedure are consistent with our theory, both for intact and hippocampal-lesioned animals. Normal intact animals show "sensory preconditioning": ⟨AB −⟩ preexposure followed by ⟨A +⟩ training leads to transfer of association to B (Thompson, 1972); hippocampal-lesioned animals show no such transfer (Port and Patterson, 1984).

Compound preconditioning

Because redundancy compression involves reducing the amount of information in a representation, it can also be shown to impair learning if stimuli that were previously compressed are subsequently made relevant. For example, suppose ⟨AB −⟩ preexposure is followed by ⟨A +, B −⟩ training. Now it will be critical to discriminate A from B. Discrimination will proceed more slowly in the preexposed system than in a system with no ⟨AB −⟩ preexposure, as shown in Figure 16. This phenomenon, a variation on latent inhibition, is also observed in intact animals (Lubow et al., 1976). Once more, the lesioned model shows no effects of preexposure. We know of no relevant data on compound preexposure with hippocampal-lesioned animals. Thus, our theory makes a novel prediction that hippocampal-damaged animals will not show a discrimination impairment following compound cue preexposure.

Comparison to standard feedforward network

To better understand the role of redundancy compression in our model, it is useful to compare our intact cortico-hippocampal model with the abilities of a generic, multilayer, feedforward prediction network (Rumelhart and McClelland, 1986) trained to map from CS cues to an expectation of the US. Like our cortico-hippocampal model, such a prediction network can be trained to learn rerepresentations of stimuli but does so within an architecture more like our cortical-only network (Fig. 17). Because this generic feedforward network does not perform autoencoding, it is not biased by stimulus-stimulus redundancy compression. Accordingly, we do not expect it to show latent inhibition. Simulations of the network in Figure 18 confirm this expectation. Although this prediction network uses the same powerful multilayer learning algorithm as our hippocampal model to firm stimulus rerepresentations, it is only trained to map from stimulus inputs (CSs and contextual cues) to a prediction of the US. Without the additional sensitivity to all stimulus-stimulus (e.g., CS-CS) relationships embodied by our hippocampal network, this generic feedforward network does not show redundancy compression.

COMPARISON TO OTHER THEORIES OF HIPPOCAMPAL FUNCTION

As noted earlier, several different representational interpretations of hippocampal region function have previously been proposed, including stimulus selection, chunking, cue con-

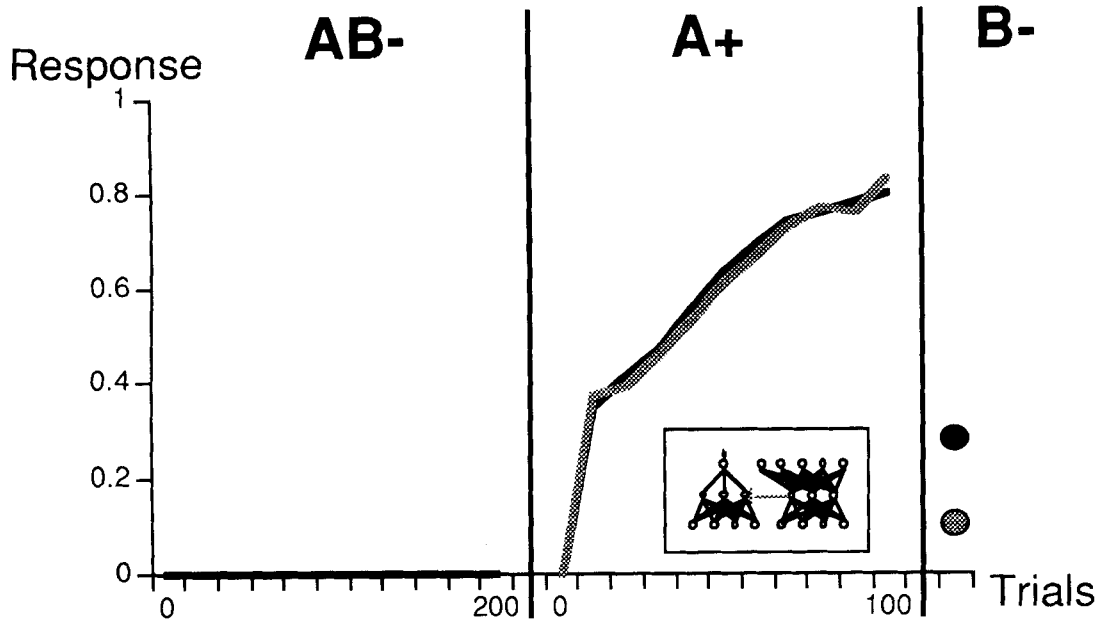


Fig. 15. Sensory preconditioning in the intact model. An ⟨AB−⟩ preexposure phase results in redundancy compression combining the representations of co-occurring stimuli A and B. During a second phase of ⟨A+⟩ training, some of the association will transfer to B. In a third testing phase, there will be some response to B (black line). As a control condition, ⟨A+⟩ training with no ⟨AB−⟩ preexposure (gray line) results in little or no response to B. Simulation results are consistent with the empirical findings of Port and Patterson (1984).

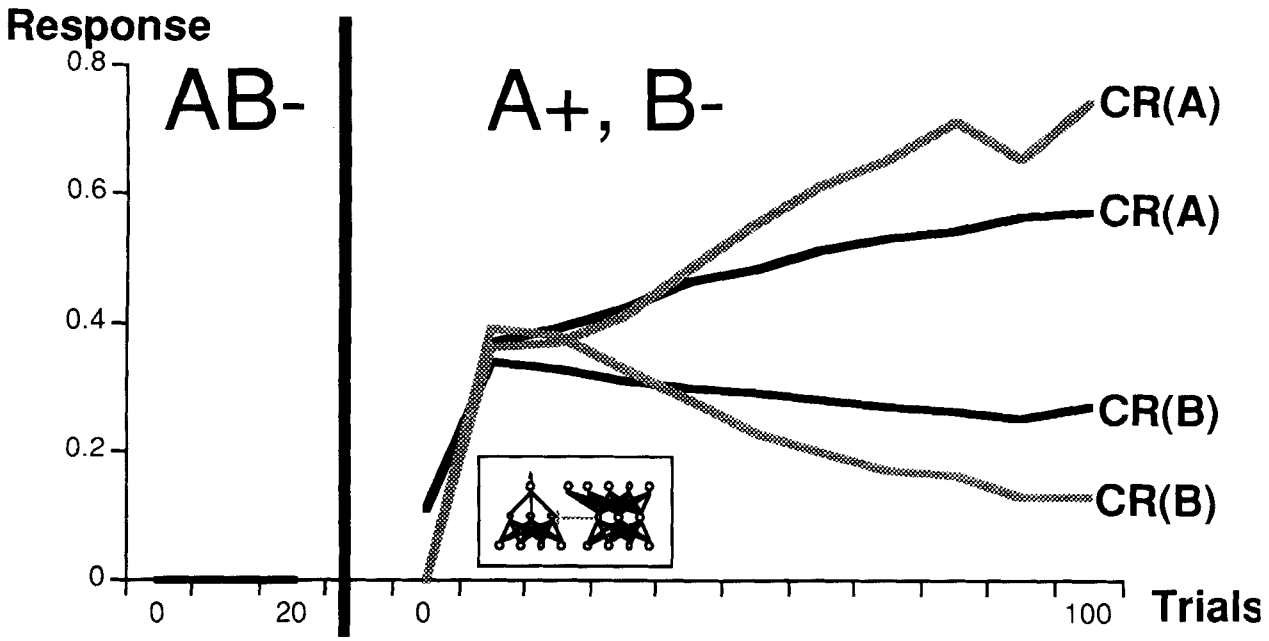


Fig. 16. Compound preexposure in the intact model. As in sensory preconditioning, an ⟨AB−⟩ preexposure phase moves the representations of co-occurring stimuli A and B closer together. If the second phase requires discriminating ⟨A+, B−⟩, the preexposure will impair this learning (black line) relative to the control condition of no preexposure (gray line). Simulation results are consistent with empirical findings of Lubow et al. (1976).

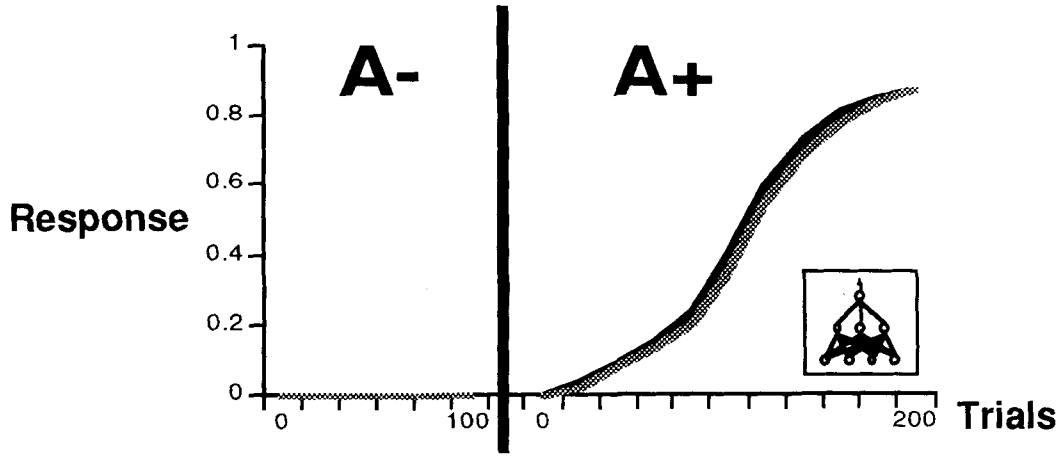


Fig. 17. A generic feedforward multilayer network (Rumelhart et al., 1986) uses a multilayer learning algorithm to modify representations in its internal layer of nodes. This network can be trained to map from stimulus inputs to a prediction of the outcome. However, unlike a predictive autoencoder (i.e., our hippocampal model) that is forced to also reproduce its inputs, this network does not perform stimulus-stimulus learning. Thus, the generic feedforward network will not show such effects as latent inhibition that depend on stimulus-stimulus redundancy compression.

figuration, and contextual coding. We now review these in more detail and show how our theory can be viewed as a computational mechanism that derives these prior qualitative characterizations of hippocampal function as task-specific special cases.

Stimulus selection

Early theorists viewed the hippocampal region as an attentional control mechanism that alters stimulus selection through a process of inhibiting attentional responses to stimuli that are nonsignificant (Grastyan et al., 1959), not correlated with reinforcement (Douglas and Pribram, 1966; Douglas, 1972; Kimble, 1968), or irrelevant in the prediction of reinforcement (Moore, 1979; Solomon, 1979). These are precisely the training conditions under which our theory expects the hippocampal region to perform redundancy compression, minimizing the representational resources allocated to these cues.

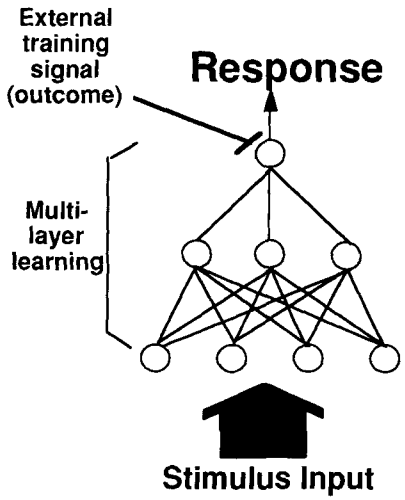


Fig. 18. Absence of latent inhibition in the generic feedforward multilayer network of Figure 17. There is no stimulus-stimulus redundancy compression in the ⟨A−⟩ preexposure phase, and so ⟨A+⟩ training (black line) is not impaired relative to a control condition with no ⟨A−⟩ preexposure (gray line).

Considering the theoretical literature on classical conditioning behavior, we note that psychological theories of stimulus selection in conditioning have traditionally fallen into two broad classes. *Reinforcement modulation theories* of stimulus selection (e.g., Rescorla and Wagner, 1972) consider the reinforcing value of the US to be modulated by the degree to which the US is unexpected given all cues present on a trial. Reinforcement modulation theories can explain effects such as conditioned inhibition (Marchant et al., 1972), that depend on cue competition, but they cannot explain other stimulus selection effects such as latent inhibition (Lubow, 1973) and reversal facilitation (see Sutherland and Mackintosh, 1971) that do not. In contrast, *sensory modulation theories* of stimulus selection (e.g., Mackintosh, 1975; Pearce and Hall, 1980) focus on modulation of CSs via mechanisms of differential attention or salience for cues that are better predictors of the US. These theories can explain latent inhibition and reversal facilitation but not conditioned inhibition. Still other stimulus selection effects, such as blocking (Kamin, 1969) and overshadowing (Kehoe, 1981), can be explained both in terms of reinforcement modulation and in terms of sensory modulation. Thus, within the theoretical and experimental literature on stimulus selection in classical conditioning, we can identify three categories of stimulus selection phenomena: those that can be uniquely explained by reinforcement modulation, those that can be uniquely explained by sensory modulation, and those that can be explained by both.

Our computational theory incorporates instantiations of both types of stimulus selection. The cortical (cerebellar) model adjusts its associations according to the error-correcting principle of the Rescorla-Wagner (1972) rule and, thereby,

instantiates stimulus selection through reinforcement modulation. Meanwhile, the sensory recoding that occurs in our hippocampal model instantiates a form of sensory modulation. Thus, our theory predicts that hippocampal lesions should disrupt stimulus selection phenomena that depend on sensory modulation but not those that depend on reinforcement modulation (some simulations shown in previous section).

A review of the available empirical data is consistent with this division of stimulus selection behaviors. For example, hippocampal lesions disrupt sensory modulation effects such as reversal facilitation (Berger and Orr, 1983) and latent inhibition (Solomon and Moore, 1975), but they do not impair reinforcement modulation effects such as conditioned inhibition (Solomon, 1977). Stimulus selection effects that are explained by both sensory and reinforcement modulation (such as blocking and overshadowing) are predicted in our model to result from combined stimulus selection mechanisms in the hippocampal and cortical regions. Thus, our theory expects that following hippocampal damage there may either be some diminution to these phenomena or else the remaining mechanisms of reinforcement modulation may suffice to generate the behaviors. More specifically, our hippocampal-lesioned model will either show or fail to show these behaviors depending on various learning parameters. These indeterminate, parameter-dependent, implications of our theory are consistent with the contradictory empirical results that have been reported for these phenomena. For example, early data suggested that blocking depends on an intact hippocampus (Solomon, 1977), while more recent studies have had difficulty replicating this result (Garrud et al., 1984).

Recently, Schmajuk and DiCarlo (1991, 1992) have extended psychological (behavioral) theories of stimulus selection with a computational model of hippocampal processing in conditioning. They have proposed a connectionist network model based on the idea that the hippocampus is essential for stimulus selection through the computation of an aggregate error in the animal's expectation of the US. The Schmajuk-DiCarlo model can account for many conditioning data, but it also fails to address several hippocampal-dependent phenomena, such as stimulus-stimulus effects. (In contrast, our theory does account for hippocampal dependence of such stimulus-stimulus effects as latent inhibition and sensory preconditioning.) The Schmajuk-DiCarlo model also makes some erroneous predictions. For example, their model predicts a hippocampal-lesion deficit for conditioned inhibition; our theory expects no such deficit. Empirical studies have found no deficit for conditioned inhibition after hippocampal lesion (Solomon, 1977). The Schmajuk-DiCarlo model also makes a strong prediction that the hippocampus plays an essential role in stimulus-selection behaviors such as blocking. As noted above, empirical data on the role of the hippocampus in blocking is mixed. The assumption that the hippocampus is critical for blocking is also at odds with data and theory that argue for a sufficient cerebellar circuit for blocking and overshadowing in motor-reflex conditioning (see Donegan et al., 1989; Gluck et al., 1993). In contrast to the implications of the Schmajuk-DiCarlo model, the aforementioned anatomical and behavioral data seem to support the idea, made explicit in our cortico-hippocampal theory, that the hippocampus is one, but not the only, substrate of stimulus selection in associative learning.

Chunking and cue configuration

The term *chunking* has traditionally referred to a learning process in which a set of stimuli, concepts, or features comes to be treated as a unary whole or "chunk" (Miller, 1956; Estes, 1972). Wickelgren (1979) proposed that the hippocampus participates in this chunking process. More recently, Wickelgren's chunking idea has been extended and elaborated by Sutherland and Rudy (1989) who propose that the hippocampus provides the neural basis for the acquisition and storage of configural events, while other brain systems store only direct cue-outcome associations. This is consistent with their data demonstrating that some configural tasks such as negative patterning ($\langle A+, B+, AB- \rangle$) are especially sensitive to hippocampal damage (e.g., Rudy and Sutherland, 1989).

In contrast to Sutherland and Rudy's characterization of hippocampal function, the role of the hippocampal region in our theory is not solely, or even primarily, one of configural association. Rather, through a process of stimulus recoding, the hippocampal region may be required to form configural associations depending on the constraints of the training environment and the status of the initial stimulus representations. Frequent cue configurations will be chunked through redundancy compression and can be discriminated from their components through predictive differentiation. Configural learning tasks such as negative patterning must often be solved in this way, and thus we expect that there will often be a lesioned deficit with configural learning.

It is important to note, however, that without the hippocampal network, the cerebellar and cortical networks in our model are left with a (fixed) set of weights in the lower layer, which perform a (static) recoding of stimulus inputs. Depending on the initial (and fixed) stimulus representation in this hippocampal-lesioned network, there is always some probability that these representations will be sufficient to allow the cortical and cerebellar networks to learn other configural associations. By appropriate choice of initial parameters in our model, it is possible to construct lesioned systems that can, on average, solve the negative patterning problem.

Thus, while our theory is similar in spirit to the implications of Sutherland and Rudy's proposal, our computational model expects only a general tendency for a hippocampal role in configural tasks, depending on the initial status of the stimulus representations. Recent experimental studies have likewise shown that hippocampal-lesioned animals can also solve some configural tasks (e.g., Whishaw and Tomie, 1991; Gallagher and Holland, 1992). This points to a general problem with configurality as the basis of a predictive theory. Whether a task is "configural" depends critically on the stimulus representation that the animal or network uses to find a solution. Except in the most transparent of training tasks with simple analyzable stimuli, this underlying representation may be unknown.

Context sensitivity

A recurring theme in interpretations of hippocampal function is its role in providing a "contextual tag" for associative learning (Hirsh, 1974; Nadel and Willner, 1980; Winocur et al., 1987; Penick and Solomon, 1991). In analyses of context and conditioning, the context is usually considered to be a

set of background cues that are present in the experimental setup but distinct from experimentally manipulated conditioned and unconditioned stimuli. These contextual stimuli might include visual features of the room, background noises, temperature, and so on.

Some researchers have viewed contextual stimuli and conditioned stimuli as being functionally equivalent (e.g., Rescorla and Wagner, 1972; Mackintosh, 1975). Because contextual stimuli are uniformly present, they receive reinforced presentation whenever the US occurs and nonreinforced presentation at all other times. Therefore, the constant contextual stimuli are usually nonreliable predictors of sparse US arrival and thus tend to acquire little association—that is, they are "tuned out" through stimulus selection. A second view of context is that contextual stimuli provide an environment that defines, colors, and contributes to the representation of the CSs (Nadel and Willner, 1980; Balsam and Tomie, 1985). In this view, what is context depends on what is foreground; thus in one situation, the color of the room, if it predicts the US, may be a foreground cue, whereas in other situations it may simply be one feature of a background environment.

In our cortico-hippocampal theory, the presence of all cues (including contextual cues) affects the representation of all other cues. No class of cue has special meaning except as it is a reliable and useful predictor of reinforcement. At the same time, stimulus recoding in the hippocampal network, by virtue of its sensitivity to regularities in stimulus-stimulus relationships, is especially influenced by the context in which cues have been presented. Thus, the presence or absence of any contextual cue can affect the final internal representation. In contrast, the lesioned model does not recode stimuli based on stimulus-stimulus regularities and thus is not nearly so sensitive to contextual effects.

To illustrate the intact model's sensitivity to context, consider for example, the following training procedure. An animal first undergoes ⟨A+⟩ training in one environment (context). Our theory expects an internal representation to be constructed that focuses on A but includes information about all contextual cues as well. This internal representation is then associated with a prediction of the US. If A is then presented in a different context, the internal representation of A will differ slightly from the one active during training. This new internal representation will probably evoke some of the US prediction, but only in as much as it recalls the representation active during training. Therefore, the model predicts a decrease in responding to A in the new context, as shown in Figure 19. In the lesioned model, however, a static internal

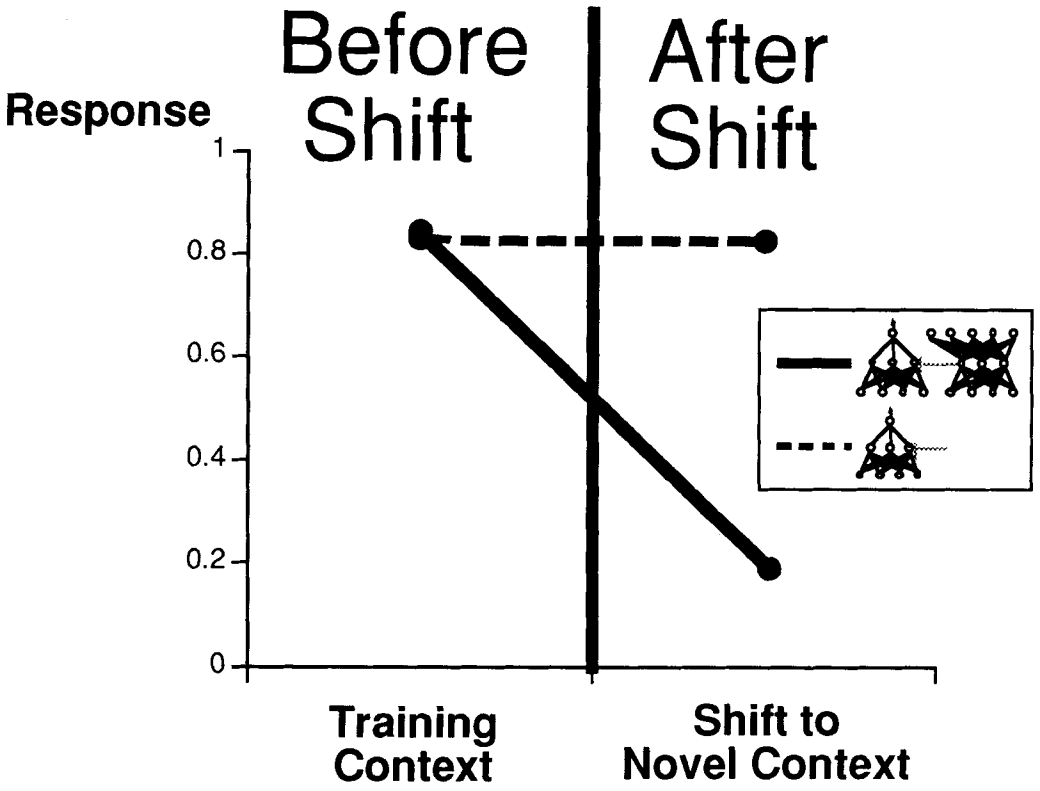


Fig. 19. Contextual sensitivity. After ⟨A+⟩ training in one context, both the intact (solid line) and lesioned models (dashed line) yield a predicted strong response to A. When the context is shifted, the lesioned model continues to respond strongly to A. However, the intact model shows a much weaker response. In the intact model, stimulus-stimulus learning ensures that no cue (even the constant contextual cues) is completely ignored. Presentation of A in the new context will therefore evoke an internal representation that differs to some degree from the learned representation, and response decrements accordingly. In the lesioned model, with no stimulus-stimulus learning, background contextual cues are simply "tuned out" of the representation, and changing them has little effect on the response to A. Simulation results are consistent with the empirical findings of Penick and Solomon (1991).

recoding means that contextual cues are simply tuned out during ⟨A+⟩ learning, and no decrease in response is expected in the new context. The expected ability of our lesioned model to tune out irrelevant contextual cues is an example of reinforcement-modulated selective attention that we expect to be retained in hippocampal-lesioned animals. Consistent with our expectations, empirical studies have shown that normal animals decrease responding to A after context shift while hippocampal lesioned animals do not (Penick and Solomon, 1991).

When is the hippocampus involved?

The earlier characterizations of hippocampal function reviewed above focused primarily on differentiating between tasks that do—and do not—require an intact hippocampus. This approach has yielded a diverse and important empirical data base that constrains all theories of hippocampal function. Some researchers have interpreted these data as implying that the hippocampus is not involved in behaviors that survive hippocampal lesion—such as acquisition ⟨A+⟩ and stimulus discrimination ⟨A+, B−⟩.

In contrast, our theory suggests that the hippocampal region of an intact animal is involved in learning even the most elementary associations (see also Eichenbaum et al., 1992, and McNaughton et al., 1992, for related interpretations). Most of the empirical data analyzed in this paper have been from learning paradigms that are variations and extensions of elementary acquisition and discrimination training. Although hippocampal damage may not always be clearly evident from analyses of error rates and trials-to-criterion during initial acquisition, we have interpreted behavioral measures of transfer performance and generalization as implying hippocampal involvement throughout learning. As described earlier, this approach suggests several new studies in which we expect both intact and lesioned animals to behave similarly on an initial learning task but differently on a subsequent transfer task.

This view of hippocampal involvement in even the simplest forms of acquisition suggests a possible interpretation of some seemingly paradoxical results. One example, described earlier, is the often-reported facilitation of simple discrimination learning in hippocampal lesioned animals (see Fig. 8). A related result is the finding that disruption of hippocampal activity during conditioning—such as from inducing seizures with injections of penicillin, medial septal lesions, or electrical stimulation—impairs learning more than removal of the hippocampus altogether (Solomon et al., 1983). At first glance, these two findings seem contradictory. If removal of the hippocampus facilitates standard acquisition learning, why should disrupting the hippocampus impair learning?

Hippocampal disruption might be approximated in our model by adding random noise to the activation levels of internal units in the hippocampal network. This will disrupt the representational "teaching signals" sent to the cortical network, and thereby cause the cortical network to develop an internal representation that is continually and randomly changing. In contrast, our model of a hippocampal-lesioned animal presumes that the cortical representations remain fixed throughout learning due to the absence of any representational teaching signals. We expect that learning with such a fixed stimulus representation should be faster (and more complete) than trying to learn with a randomly changing representation. As expected, the simulation results shown in Figure 20 are broadly consistent with the empirical results reported by Solomon et al. (1983): learning in the disrupted system is severely impaired—slower, in fact, than in either the intact or lesioned systems.

Thus, these seemingly paradoxical results—that discrimination learning may be facilitated after hippocampal lesion, while a disrupted hippocampus is worse than learning with

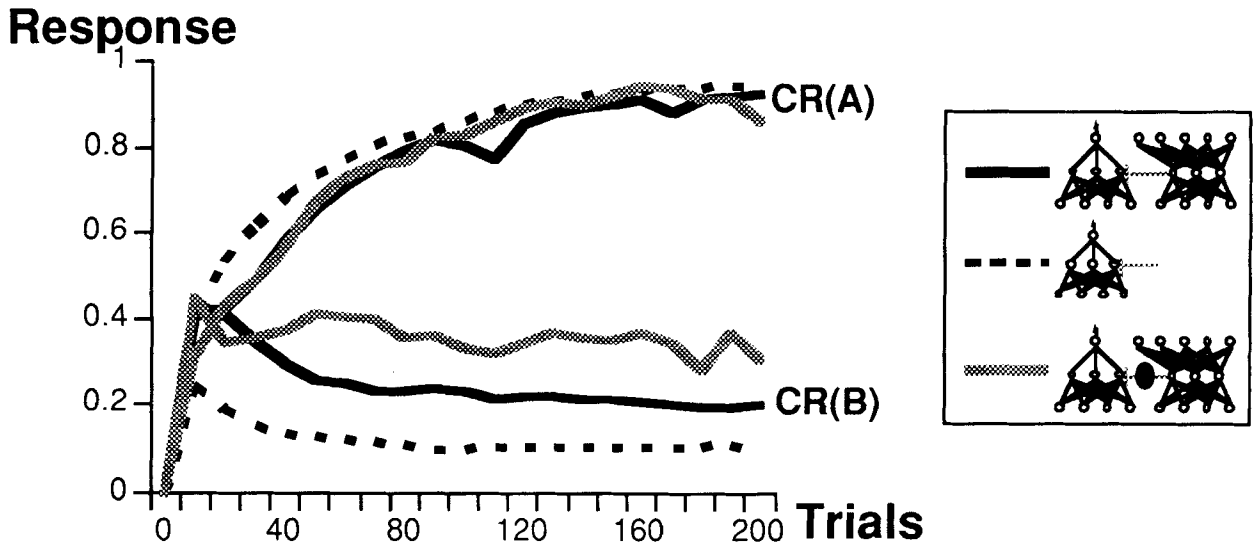


Fig. 20. Disruption of the hippocampal model during ⟨A+⟩ learning (gray line) impairs learning more than either an intact model (black line) or a lesioned model (dashed line). Disruptions are modeled by randomizing the training signal from the hippocampal model internal representation to the cortical model internal layer. Simulation results are consistent with the empirical findings of Solomon et al. (1983).

no hippocampus at all—can be reconciled within our theory. These results provide further evidence that the hippocampal region is always involved in discrimination learning, even in the simplest tasks such as stimulus discrimination or acquisition.

DISCUSSION

This paper has presented a theory of cortico-hippocampal interaction in discrimination learning. Our basic thesis is that one (but not necessarily the only) function of the hippocampal region is the recoding of internal stimulus representations to facilitate learning. These new representations are expected to enhance the discriminability of differentially predictive cues while compressing the representation of redundant cues. Other brain regions, including the cerebral and cerebellar cortices, are presumed to use these hippocampal representations to recode their own stimulus representations. In the absence of an intact hippocampal region, these other brain regions can still acquire some associative learning tasks using previously established fixed representations. However, this ahippocampal learning is expected to show consistent and predictable differences from normal learning, especially inappropriate generalization to novel task demands (see also Eichenbaum et al., 1992).

The theory, instantiated as a connectionist network model, provides a simple and unified interpretation of the trial-level effects of hippocampal lesions on classical conditioning. Applying the connectionist network model to conditioning tasks demonstrates how the theory accounts for the effects of hippocampal lesions on simple discrimination, reversal learning, stimulus generalization, latent inhibition, sensory preconditioning, and contextual sensitivity. These analyses suggest that the effects of hippocampal damage may be especially informative in studies of two-phase transfer tasks where the effects of altered stimulus representations are most apparent. This suggests that a profitable direction for future empirical studies may be to examine other relatively simple two-phase learning paradigms, in which both intact and hippocampal-lesioned animals behave similarly on an initial learning task but respond differently on subsequent transfer and generalization tasks. For example, the theory makes novel predictions regarding the effects of easy-hard transfer and compound preexposure.

Temporal processing

The focus of this paper has been on the role of the hippocampal region in classical conditioning. Because the model, in its current form, addresses only trial-level effects, we have not addressed the substantial body of data documenting the importance of the hippocampal region for temporal processing aspects of this paradigm. For example, hippocampal-lesioned animals are impaired at long ISI conditioning (Akase et al., 1989) and trace conditioning (Moyer et al., 1990). In previous theoretical analyses we have shown how a recurrent network model of the cerebellum can account for the temporal characteristics of CR topography during classical conditioning (Gluck et al., 1993). As one would expect from a model of the cerebellum alone, this real-time model can account for temporal behaviors exhibited by hippocampal-damaged animals but not for behaviors requiring an intact hippocampus.

Exploring the predictions of our theory with respect to temporal behaviors would require an extension of the cortico-hippocampal model presented here. Presumably this might involve including a recurrent network in place of the present trial-level cortical network. The key issue for such future analyses is whether or not the same representational recodings that we have postulated in this theory can be generalized and extended to account for hippocampal-lesion deficits in temporal processing. A unification of the representational and temporal processing functions of the hippocampal region would represent an important step forward in our understanding of hippocampal-region function in associative learning.

Spatial navigation

Hippocampal-lesioned animals are also significantly impaired on spatial tasks such as navigation through a milky pool to a submerged platform (Gallagher and Holland, 1992; Morris, 1983). As a result of these and similar experiments, several researchers have proposed that the hippocampus is involved in the construction of spatial maps (e.g., O'Keefe and Nadel, 1978). Several researchers have argued that the high internal recurrency within the hippocampus could allow it to serve as an autoassociator, mapping from partial information about the current environment to a more complete representation (e.g., McNaughton, 1989; Rolls, 1990). One possibility is to define spatial maps as composed of sets of complex configural associations representing places (McNaughton, 1989; McNaughton and Nadel, 1990). In one place, there may be many views, depending on which way the animal is facing, the location of landmarks, etc. The hippocampal autoassociator would be able to map from one of these views to the full representation of the current place. With this interpretation, place learning need not be fundamentally different from any other kind of representational learning. However, because of the need for such complex representations in spatial tasks, these behaviors might be especially sensitive to hippocampal damage. The predictive autoencoder of our hippocampal network is a more powerful version of these simple autoassociators, and thus our model is potentially consistent with this line of reasoning. A more complete theory and model of spatial navigation, and the representational demands it places on an animal, would be needed before these ideas could be developed in a more rigorous fashion.

Cognitive processes and human memory

It is clear that a considerable amount of further work is required before the ideas presented in this paper can be applied to a wider range of animal learning behaviors including temporal processing and spatial navigation. Mapping from our theory of hippocampal function in animal conditioning to an understanding of the hippocampal region in human memory would be an even larger leap. Mechanistic models for human learning and memory, and the representational demands they place on the learner, are generally not nearly as well characterized for higher cognitive forms of learning as they are for simpler associative tasks such as conditioning. Models and representational demands must be specified for these cognitive tasks before the current approach can be applied to comparative studies of normal and hippocampal-damaged (amne-

sic) humans. We note, however, that research on behavioral correspondences between classical conditioning and human category learning (e.g., Gluck and Bower, 1988a, 1988b; Gluck et al., 1989; Gluck, 1991) may suggest possible avenues for integrating animal and human models of hippocampal-region function.

ACKNOWLEDGMENTS

For their thoughtful comments and advice on this work, the authors are indebted to Terry Allard, David Amaral, Carol Barnes, Gyorgy Buzsaki, Howard Eichenbaum, Bill Estes, Paul Glauthier, Ofer Goren, Richard Granger, Bill Hirst, Barbara Knowlton, James McGaugh, Bruce McNaughton, Lynn Nadel, Jerry Rudy, Daniel Schacter, Paul Solomon, Larry Squire, Herb Terrace, and Richard Thompson. The assistance of Adrianna Herrera and Pratiksha Pandit is also gratefully acknowledged. This research was supported by the Office of Naval Research through the Young Investigator Program and by grant N00014-88-K-0112.

References

Akase E, Alkon DL, Disterhoft JF (1989) Hippocampal lesions impair memory of short-delay conditioned eye blink in rabbits. Behav Neurosci 103:935–943.

Anderson JA (1977) Neural models with cognitive implications. In: Basic processes in reading: perception and comprehension (LaBerge DL, Samuels SJ, eds), pp 27–90. Hillsdale, NJ: Erlbaum.

Balsam P, Tomie A (1985) Context and conditioning. Hillsdale, NJ: Erlbaum.

Berger TW, Orr WB (1983) Hippocampectomy selectively disrupts discrimination reversal learning of the rabbit nictitating membrane response. Behav Brain Res 8:49–68.

Brown T, Kairiss E, Keenan C (1990) Hebbian synapses: biophysical mechanisms and algorithms. Ann Rev Neurosci 13:475–511.

Buszaki G (1989) Two-stage model of memory trace formation: a role for "noisy" brain states. Neuroscience 31:551–570.

Donegan NH, Gluck MA, Thompson RF (1989) Integrating behavioral and biological models of classical conditioning. In: Computational models of learning in simple neural systems. Vol. 22. Psychology of Learning and Motivation (Hawkins RD, Bower GH, eds), pp 109–156. New York: Academic Press.

Douglas R (1972) Pavlovian conditioning and the brain. In: Inhibition and learning (Boakes RA, Halliday MS, eds), pp 529–549. London: Academic Press.

Douglas R, Pribram K (1966) Learning and limbic lesions. Neuropsychologia 4:192–220.

Eichenbaum H, Buckingham J (1991). Studies on hippocampal processing: experiment, theory, and model. In: Neurocomputation and learning: foundations of adaptive networks (Gabriel M, Moore J, eds), pp 171–231. Cambridge, MA: MIT Press.

Eichenbaum H, Cohen N, Otto T, Wible C (1992) Memory representation in the hippocampus: functional domain and functional organization. In: Memory: organization and locus of change. (Squire L, Lynch G, Weinberger N, McGaugh J, eds), pp 163–204. Oxford: Oxford University Press.

Eichenbaum H, Fagan A, Cohen N (1986) Normal olfactory discrimination learning set and facilitation of reversal learning after medial-temporal damage in rats: implication for an account of preserved learning abilities in amnesia. J Neurosci 6:1876–1884.

Eichenbaum H, Fagan A, Mathews P & Cohen N (1988) Hippocampal system dysfunction and odor discrimination learning in rats: impairment or facilitation depending on representational demands. Behav Neurosci 102:331–339.

Eichenbaum H, Otto T, Cohen N (1992) The hippocampus: what does it do? Behav Neural Biol 57:2–36.

Eichenbaum H, Otto T, Wible C, Piper J (1991) Building a model of the hippocampus in olfaction and memory. In: Olfaction as a model for computational neuroscience (Davis J, Eichenbaum H, eds), pp 167–210. Cambridge, MA: MIT Press.

Estes W (1972) An associative basis for coding and organization in memory. In: Coding processes in human memory (Melton A, Martin E, eds), pp 161–190. Washington, DC: VH Winston.

Gallagher M, Holland P (1992) Preserved configural learning and spatial learning impairment in rats with hippocampal lesions. Hippocampus 2:81–88.

Garrud P, Rawlins JNP, Mackintosh NJ, Goodal G, Cotton MM, Feldon J (1984) Successful overshadowing and blocking in hippocampectomized rats. Behav Brain Res 12:39–53.

Gluck MA (1991) Stimulus generalization and representation in adaptive network models of category learning. Psychol Sci 2:1–6.

Gluck MA, Bower, GH (1988a) From conditioning to category learning: an adaptive network model. J Exp Psychol [Gen] 117:225–244.

Gluck MA, Bower GH (1988b) Evaluating an adaptive network model of human learning. J Mem Lang 27:166–195.

Gluck MA, Bower GH (1990) Component and pattern information in adaptive networks. J Exp Psychol [Gen] 119:105–109.

Gluck MA, Granger R (1993) Computational models of the neural bases of learning and memory. Annu Rev Neurosci 16:667–706.

Gluck MA, Bower GH, Hee MR (1989) A configural-cue network model of animal and human associative learning. pp. 323–332. 11th Annual Conference of the Cognitive Science Society, Ann Arbor, MI.

Gluck MA, Goren O, Myers C, Thompson RF (1993) A higher-order recurrent network model of the cerebellar substrates of response timing in motor-reflex conditioning. J Cogn Neurosci (in press).

Gormezano I, Kehoe EK, Marshal BS (1983) Twenty years of classical conditioning research with the rabbit. Prog Psychobiol Physiol Psychol 10:197–275.

Grastyan E, Lissak K, Madarasz I, Donhoffer H (1959) Hippocampal electrical activity during the development of conditioned reflexes. Electroencephalogr Clin Neurophysiol 11:409–430.

Hinton GE (1989) Connectionist learning procedures. Artif Intell 40: 185–234.

Hirsh R (1974) The hippocampus and contextual retrieval of information from memory: a theory. Behav Biol 12:421–444.

Kamin LJ (1969) Predictability, surprise, attention, and conditioning. In: Punishment and aversive behavior (Campbell B, Church R, eds), pp 279–296. New York: Appleton-Century-Crofts.

Kehoe EJ (1981) Stimulus selection and combination in classical conditioning in the rabbit. In: Classical conditioning (Gormezano I, Prokasy WF, Thompson RF eds), pp 161–196. Hillsdale, NJ: Erlbaum.

Kimble DP (1968) Hippocampus and internal inhibition. Psychol Bull 70:285–295.

Kohonen T (1984) Self-organization and associative memory. New York: Springer-Verlag.

Lawrence D (1952) The transfer of a discrimination along a continuum. J Comp Physiol Psychol 45:511–516.

Lubow RE (1973) Latent inhibition. Psychol Bull 79:398–407.

Lubow RE, Rifkin B, Alek M (1976) The context effect: the relationship between stimulus pre-exposure and environmental pre-exposure determines subsequent learning. J Exp Psychol [Anim Behav] 2:38–47.

Mackintosh NJ (1975) A theory of attention: variations in the associability of stimuli with reinforcement. Psychol Rev 82:276–298.

Mackintosh NJ (1983) Conditioning and associative learning. Oxford: Oxford University Press.

Mackintosh NJ, Little L (1970) An analysis of transfer along a continuum. Can J Psychol 24:362–369.

Marchant HG, Mis FW, Moore JW (1972) Conditioned inhibition of the rabbit's nictitating membrane response. J Exp Psychol Gen 95: 408–411.

McNaughton BL (1989) Neuronal mechanisms for spatial computation and information storage. In: Neural connections, mental computations (Nadel L, Cover LA, Culicover P, Harnish RM, eds), pp 285–350. Cambridge, MA: MIT Press.

McNaughton BL, Nadel L (1990) Hebb-Marr networks and the neurobiological representation of action in space. In: Neuroscience and connectionist theory (Gluck MA, Rumelhart DE, eds), pp 1–63. Hillsdale, NJ: Erlbaum.

McNaughton B, Leonard B, Chen L (1992) Cortical-hippocampal interaction and cognitive mapping: an hypothesis based on reintegration of the parietal and inferotemporal pathways for visual processing. Psychobiology 236–246.

Miller G (1956) The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychol Rev 63: 81–97.

Mishkin M (1982) A memory system in the monkey. Philos Trans R Soc Lond [Biol] 298:85–92.

Moore JA (1979) Brain processes and conditioning. In: Mechanisms of learning and motivation (Dickinson RA, Boakes RA, eds), pp 111–142. Hillsdale, NJ: Erlbaum.

Morris R (1983) An attempt to dissociate "spatial-mapping" and "working-memory" theories of hippocampal function. In: Neurobiology of the hippocampus (Seifert W, ed), pp 405–432. London: Academic Press.

Morris RGM, Garrud P, Rawlins JNP & O'Keefe J (1982) Place navigation impaired in rats with hippocampal lesions. Nature 297: 681–683.

Moyer JR, Deyo RA, Disterhoft JF (1990) Hippocampectomy disrupts trace eye-blink conditioning in rabbits. Behav Neurosci 104: 243–252.

Nadel L, Willner J (1980) Context and conditioning: a place for space. Physiol Psychol 8:218–228.

Nosofsky RM (1984) Choice, similarity, and the context theory of classification. J Exp Psychol [Learn Mem Cogn] 10:104–114.

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

Parker D (1985) Learning logic. Cambridge, MA: Center for Computational Research in Economics and Management Science, MIT.

Pearce JM, Hall G (1980) A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. Psychol Rev 87:532–552.

Penick S, Solomon R (1991) Hippocampus, context and conditioning. Behav Neurosci 105:611–617.

Port R, Patterson MM (1984) Fimbrial lesions and sensory preconditioning. Behav Neurosci 98:584–589.

Port R, Mikhail A, Patterson M (1985) Differential effects of hippocampectomy on classically conditioned rabbit nictitating membrane response related to interstimulus interval. Behav Neurosci 99:200–208.

Port R, Romano A, Patterson M (1986) Stimulus duration discrimination in the rabbit: effects of hippocampectomy on discrimination and reversal learning. Physiol Psychol 14:124–129.

Reiss S, Wagner AR (1972) CS habituation produces a "latent inhibition" effect but no "conditioned inhibition." Learn Motiv 3: 237–245.

Rescorla RA, Wagner AR (1972) A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and non-reinforcement. In: Classical conditioning II: current research and theory (Black AH, Prokasy WF, eds), pp 64–99. New York: Appleton-Century-Crofts.

Riley D (1968) Discrimination learning. Boston: Allyn & Bacon.

Rolls E (1990) Theoretical and neurophysiological analysis of the functions of the primate hippocampus in memory. Cold Spring Harb Symp Quan Biol 55:995–1006.

Rudy JW, Sutherland RJ (1989) The hippocampal formation is necessary for rats to learn and remember configural discriminations. Behav Brain Res 34:97–109.

Rumelhart DE, Hinton GE, Williams RJ (1986) Learning internal representations by error propagation. In: Parallel distributed processing: explorations in the microstructure of cognition. Vol. 1. Foundations (Rumelhart D, McClelland J, eds), pp 318–362. Cambridge, MA: MIT Press.

Rumelhart DE, McClelland JL (1986) Parallel distributed processing: explorations in the microstructure of cognition. Vol. 1. Foundations. Cambridge, MA: MIT Press.

Schmajuk NA, DiCarlo JJ (1991) Neural dynamics of hippocampal modulation of classical conditioning. In: Neural network models of conditioning and action (Grossberg CNS, Staddon JER, eds), pp 149–180. Hillsdale, NJ: Erlbaum.

Schmajuk NA, DiCarlo JJ (1992) Stimulus configuration, classical conditioning and hippocampal function. Psychol Rev 99:268–305.

Schmajuk NA, Moore JW (1985) Real-time attentional models for classical conditioning and the hippocampus. Physiol Psych 13: 278–290.

Schmaltz LW, Theios J (1972) Acquisition and extinction of a classically conditioned response in hippocampectomized rabbits (*Oryctolagus cuniculus*). J Comp Physiol Psychol 79:328–333.

Scoville WB, Milner B (1957) Loss of recent memory after bilateral hippocampal lesions. J Neurol Neurosurg Psychiat 20:11–21.

Shepard RN (1958) Stimulus and response generalization: deduction of the generalization gradient from a trace model. Psychol Rev 65: 242–256.

Shepard RN (1987) Towards a universal law of generalization for psychological science. Science 237:1317–1323.

Solomon PR (1977) Role of the hippocampus in blocking and conditioned inhibition of rabbit's nictitating membrane response. J Comp Physiol Psychol 91:407–417.

Solomon PR (1979) Temporal versus spatial information processing theories of hippocampal function. Psychol Bull 86:1272–1279.

Solomon PR, Moore JW (1975) Latent inhibition and stimulus generalization of the classically conditioned nictitating membrane response in rabbits (*Oryctolagus cuniculus*) following dorsal hippocampal ablation. J Comp Physiol Psychol 89:1192–1203.

Solomon P, Solomon S, Van der Schaaf E, Perry H (1983) Altered activity in the hippocampus is more detrimental to classical conditioning than removing the structure. Science 220:329–331.

Squire LR (1987) Memory and brain. New York: Oxford University Press.

Squire LR, Frambach M (1990) Cognitive skill learning in amnesia. Psychobiology 18:109–117.

Squire LR, Zola-Morgan S (1983) The neurology of memory: the case for correspondence between the findings for man and non-human primate. In: The physiological basis of memory (Deutsch JA, ed), pp 199–268. New York: Academic Press.

Stanton P, Sejnowski T (1989) Associative long-term depression in the hippocampus induced by Hebbian covariance. Nature 339: 215–218.

Sutherland NS, Mackintosh NJ (1971) Mechanisms of animal discrimination learning. New York: Academic Press.

Sutherland RJ, Rudy JW (1989) Configural association theory: the role of the hippocampal formation in learning, memory, and amnesia. Psychobiology 17:129–144.

Sutton RS, Barto AG (1981) Toward a modern theory of adaptive networks: expectation and prediction. Psychol Rev 88:135–170.

Terrace H (1963) Discrimination learning with and without "errors." J Exp Anal Behav 6:1–27.

Thompson R (1958) Primary stimulus generalization as a function of acquisition level in the cat. J Comp Physiol Psychol 51:601–606.
Thompson R (1972) Sensory preconditioning. In: Topics in learning and performance (Thompson RF, Voss JF, ed), pp 105–129. New York: Academic Press.
Thompson RF (1986) The neurobiology of learning and memory. Science 233:941–947.
Thompson RF, Gluck MA (1990) Brain substrates of basic associative learning and memory. In: Cognitive neuroscience (Weingartner HJ, Lister RF, eds), pp 24–45. New York: Oxford University Press.
Werbos P (1974) Beyond regression: new tools for prediction and analysis in the behavioral sciences. PhD Thesis, Harvard University, Boston, MA.
Whishaw I, Tomie J (1991) Acquisition and retention by hippocampal rats of simple, conditional and configural tasks using tactile and olfactory cues: implications for hippocampal function. Behav Neurosci 105:787–797.
Wickelgren WA (1979) Chunking and consolidation: a theoretical synthesis of semantic networks, configuring in conditioning, S-R versus cognitive learning, normal forgetting, the amnesic syndrome and the hippocampal arousal system. Psychol Rev 86:44–60.
Widrow B, Hoff M (1960) Adaptive switching circuits. Institute of Radio Engineers, Western Electronic Show and Convention, Convention Record 4:96–194.
Winocur G, Rawlins J, Gray JR (1987) The hippocampus and conditioning to contextual cues. Behav Neurosci 101:617–625.

APPENDIX: SIMULATION DETAILS

The input pattern on each trial consists of an 18-bit stimulus vector, specifying the values of up to 3 CSs and 15 contextual cues. Inputs are generally present (1.0) or absent (0.0), except in experiments that use intermediate values (e.g., easy-hard transfer). One learning block consists of 10 trials: there is 1 trial with each stimulus being trained, and the remainder are trials with context only (no CSs present). Contextual cues are initialized to random (0.0 or 1.0) values at the start of an experiment, and during each learning block, a random contextual cue changes value with probability .01.

Hippocampal model

The hippocampal network is a three-layer network with full connectivity between 19 input nodes and 10 internal nodes and between the internal nodes and 19 output nodes. Input consists of the 18-bit stimulus vector plus a constant 0-valued signal representing the (unknown) US value. Desired output is the same 18-bit vector plus the actual US value. Node output is calculated as:

$$y_j = f\left(\sum_i w_{ij} y_i + \theta_j\right)$$

$$f(x) = \frac{1}{(1 + e^{-x})}$$

where w_{ij} is the weight from node i to node j and θ_j is node j's bias. Weights and biases are initialized according to a uniform distribution U(−0.3, +0.3). Weight update occurs as:

$$\Delta w_{ij} \leftarrow \beta \delta_j y_i + \alpha(\Delta w_{ij})$$

$$\text{where } \delta_j = (I_j - y_j) y_j (1 - y_j) \quad \text{for output node } j$$

$$\text{and} \quad \delta_j = y_j (1 - y_j)\left(\sum_k w_{jk} \delta_k\right) \quad \text{for hidden node } j$$

On trials where the US is present, $\beta = .05$; otherwise $\beta = .005$. I_j is the desired output for node j; $\alpha = 0.9$; and θ_j is trained as if it were a weight to node j from a node that constantly output 1.0.

A trials B trials

Cortical model

The cerebellar network is a 3-layer network with full connectivity between 18 input nodes and 60 internal nodes and between the internal nodes and 1 output node. Input consists of the 18-bit pattern vector. Desired output is the actual US value. Node output is calculated as in the hippocampal network. Weights and biases are initialized according to a uniform distribution U(−0.3, +0.3), except that a random two weights from each input node are initialized from U(−3.0, +3.0). This, together with the large number of internal units, allows the lesioned model (cortical network only) to have a good range of initial weights—therefore increasing the likelihood that it can solve a random discrimination. The weight from internal layer node i to output node j is updated according to:

$$\Delta w_{ij} = \beta(\text{US} - y_j) y_i$$

On trials where the US is present, $\beta = .5$; otherwise $\beta = .05$.

There is full connection from hippocampal internal layer nodes to cortical internal nodes, with weights initialized from U(−0.3, +0.3). These weights are nonadaptive. The net error signal to cortical internal layer node j is calculated as

$$\delta_j = \sum_{h \in H} y_h w_{hj}$$

where H is the set of hippocampal network hidden units. The weights from cortical input node i to hidden layer node j are then updated as:

$$\Delta w_{ij} = \beta(\delta_j - y_j) y_i$$

Activation of the cortical network output node is interpreted as the CR.