Late-training Amnesic Deficits in Probabilistic Category Learning: A Neurocomputational Analysis

Mark A. Gluck, 1,3 Lindsay M. Oliver, 2 and Catherine E. Myers 1

1 Center for Molecular and Behavioral Neuroscience
Rutgers University
Newark, New Jersey 07102
2 Department of Psychology
Glasgow Caledonian University
Glasgow, UK

Abstract

Building upon earlier behavioral models of animal and human learning, we explore how a psychobiological model of animal conditioning can be applied to amnesic category learning. In particular, we show that the late-training deficit found in Knowlton, Squire, and Gluck’s 1994 study of amnesic category learning can be understood as a natural consequence of Gluck and Myers’s (1993) theory of hippocampal-region function, a theory that has heretofore been applied only to studies of animal learning. When applied to Knowlton et al.’s category learning task, Gluck and Myers’s model assumes that the hippocampal region induces new stimulus representations over multiple training trials that reflect stimulus–stimulus regularities in the training set. As such, the model expects an advantage for control subjects over hippocampal-damaged amnesic patients only later in training when control subjects have developed new hippocampal-dependent stimulus representations; in contrast, both groups are expected to show equivalent performance early in training. A potentially analogous early/late distinction is described for animal studies of stimulus generalization. Our analyses suggest that careful comparisons between early and late-training differences in learning may be an important factor in understanding amnesia and the neural bases of both animal and human learning.

Introduction

In their recent study of category learning in amnesia, Knowlton et al. (1994) found that the performance of amnesic patients matched that of intact subjects early in training, although their performance later in training revealed a deficit relative to controls. When applied to the same task, Gluck and Myers’s (1993) cortico-hippocampal model assumes that the hippocampal region constructs new representations of stimulus patterns. These new representations eventually may be adopted by the cortical areas that are assumed to be the sites of long-term memory and of behavioral expression. Before the hippocampal-mediated representations are constructed, the cortical regions may approximate correct behavior based on whatever preexisting representations are available. Thus, early in training, the model expects there may be little difference between an intact and hippocampal-lesioned system. Only later in training, as the hippocampal-mediated representations are available, may the intact system be able to show an advantage by using them. Thus, the model expects advantages for normal over hippocampal-lesioned subjects late—but not necessarily early—in training, just as was observed by Knowlton et al. (1994).

The training paradigm employed in the Knowlton et al. (1994) study, probabilistic category (or “classification”) learning, has played a important role in the psychology of human learning, yielding a substantial body of empirical and theoretical results (see Estes 1994). Most current for-
mal or computational models of human learning have been developed and applied to account for data from such category learning studies. One approach to modeling human category learning has been to explore how elementary behavioral principles of conditioning can be generalized within connectionist network models to account for a range of human category learning behaviors (Gluck and Bower 1988a; Estes et al. 1989; Gluck et al. 1989; Shanks 1991). In this paper, we build upon these earlier behavioral models of animal and human learning to explore how a psychobiological model of animal conditioning can be applied to amnesic category learning. The long-range goal of this approach is to understand the extent to which biological principles of learning may be conserved across humans and experimental animals.

We turn first to Gluck and Myers's (1993) computational model of cortico-hippocampal function and review data from studies of animal conditioning that support this model. We then turn to the Knowlton et al. (1994) study and show how Gluck and Myers's (1993) model accounts for the difference between early and late-training performance in amnesic patients and control subjects. We then consider the implications of this early/late distinction in the cortico-hippocampal model for other learning paradigms, such as stimulus generalization, and illustrate how the model might interpret some contradictory results that have been reported with comparative studies of intact and hippocampal-lesioned animals.

HIPPOCAMPAL FUNCTION IN ANIMAL LEARNING

The study of hippocampal-dependent memories in human amnesic patients has focused primarily on declarative memories because these are the most severely and clearly impaired in amnesia. In contrast, many simple procedural—or habit—memories survive hippocampal-region damage in both animals and humans. Animals with hippocampal-region damage show normal acquisition for classically conditioned responses to a single stimulus (e.g., Solomon and Moore 1975; Solomon 1977; Akase et al. 1989; Schmajuk et al. 1994), the ability to choose the novel stimulus from a pair of objects where one object was seen immediately before (Zola-Morgan et al. 1992), discrimination of singly presented odor stimuli in operant discrimination (Eichenbaum et al. 1988), and learning to swim to an escape platform when started from a constant location in a pool (Eichenbaum et al. 1991a). Similarly, human hippocampal-damaged amnesic patients are not impaired at acquiring a conditioned motor-reflex response (Daum et al. 1989; Woodruff-Pak 1993; Gabrieli et al. 1995), learning simple classification tasks (Knowlton et al. 1994), or learning new motor skills such as mirror drawing (Milner 1962). All of these spared tasks can be solved by incremental formation of habits or stimulus–response associations, without requiring the explicit recall of the episodic memories of individual learning sessions.

There are, however, other tasks in the animal learning literature that appear—at least superficially—to be just as procedural or habit-based as those noted above, but which are impaired after hippocampal-region damage. For example, although acquisition of a classically conditioned response is unimpaired by hippocampal-region damage, severe impairments are observed in animal conditioning tasks that require learning about unreinforced stimuli (Solomon and Moore 1975; Kaye and Pearce 1987), configurations or combinations of stimuli (Sutherland and Rudy 1989), contextual information (Hirsh 1974), or relationships that span short delays (Rawlins 1985; Port et al. 1986; Moyer et al. 1990; Zola-Morgan et al. 1992; Eichenbaum et al. 1994). These findings imply that the hippocampal-region does indeed participate in information processing during incrementally learned habit-like tasks, although this participation may not necessarily be evident from studying acquisition of an initial associative response.

A major challenge for theories of hippocampal function in animal learning is to explain the wide range of learning deficits and changes in generalization behavior seen in conditioning studies following damage to the hippocampal region. We now briefly review a recent computational theory that accounts for a wide range of these behaviors, arguing that they result from an impairment to adaptive changes in stimulus representations that normally take place in an intact animal during learning.

SIMILARITY, STIMULUS REPRESENTATION, AND HIPPOCAMPAL COMPUTATION

Gluck and Myers (1993) presented a computational theory of hippocampal function in associa-
tive learning that proposed that the hippocampal region mediates associative learning by altering the similarities among internally constructed stimulus representations. Similarity can be defined operationally as the degree to which an animal (or person) is likely to generalize what it knows about one stimulus to another stimulus (James 1896; Shepard 1958; Tversky 1977). There are many ways to conceptualize the psychological notion of similarity. Within cognitive psychology, it is common to view stimulus items as being represented as points in a multidimensional “psychological space”; within this geometric model of similarity, small interitem distances are indicative of high similarity, whereas large interitem distances are indicative of low similarity (Shepard 1957). These differences in similarity have important consequences for learning and generalization: If two stimuli are highly similar, what is learned about one will generalize strongly to the other and this will impede learning to discriminate between them. If, however, two items have low similarity (i.e., are highly dissimilar), then what is learned about one item will generalize very little to the other, and learning to discriminate between them will be relatively easy.

In several recent papers, Gluck and Myers have reviewed evidence that similarity is not fixed, but changes during learning in response to correlations between stimuli (Gluck and Myers 1993, 1995; Myers and Gluck 1994; Myers et al. 1995). Related accounts of adaptive similarity have also been proposed within cognitive psychology for theories of concept learning (Nosofsky 1989; Kruschke 1992) and categorical perception (Harnad et al. 1994; Building on the behavioral evidence for adaptive similarity in learning, Gluck and Myers (1993) described comparative behavioral analyses that suggest that animals with hippocampal-region damage are unable to alter stimulus similarity based on experience. While these hippocampal-damaged subjects can still learn whether or not to respond to an individual stimulus, they are notably impaired at many tasks involving learning relationships between stimuli, especially in the absence of explicit reinforcement.

Gluck and Myers (1993) proposed that the hippocampal region adapts stimulus similarity according to two basic principles. The first principle, termed “predictive differentiation”, is a tendency to differentiate—or reduce similarity—between stimuli that are predictive of different salient future events. For example, if stimulus A predicts reward but B does not, similarity between them should decrease, minimizing generalization between them and facilitating their discrimination. The second principle, termed “redundancy compression”, is a tendency to increase similarity between stimuli that co-occur and/or predict similar future events. Thus, if A and B reliably co-occur (and are therefore likely to make similar predictions about future reward), the similarity between A and B is increased, enhancing generalization and decreasing discriminability between A and B.

For most simple training paradigms in which there is a simple set of stimulus–stimulus correlations, the Gluck and Myers theory makes clear predictions for learning and generalization that can be deduced at a verbal or qualitative level, without computer models or simulations; in summary, the Gluck and Myers theory expects many of these changes in similarity to be evident in normal intact animals, but to be missing or altered in animals with damage to the hippocampal region. The theory thus provides a clear and formal theory for how and where to expect hippocampal-lesioned animals to differ behaviorally from normal animals. Although the theory can be applied usefully at this verbal/qualitative level, a formal instantiation of the theory allows the specification of exact predictions, as well as providing a way to explore more subtle interactions among multiple stimuli in complex training paradigms.

A CONNECTIONIST MODEL OF ADAPTIVE SIMILARITY IN CORTICO-HIPPOCAMPAL FUNCTION

This basic theory can be implemented by the connectionist network model shown in Figure 1A (Gluck and Myers 1993). The hippocampal region is modeled as a predictive autoencoder (Hinton 1989), which learns to map from stimulus inputs to outputs that reconstruct those inputs as well as predicting the correct behavioral response. To achieve this, the nodes in the internal layer must form new representations that compress redundancies in the input but preserve or differentiate information that is predictive of the behavioral response. One learning algorithm that can accomplish this is error back-propagation (Rumelhart et al. 1986), but it is important to note that this particular architecture and learning algorithm represent only one of many possible ways to instantiate this fundamental theory of representational changes. Within this modular architecture for
memory, the hippocampal region acts as a teacher, instructing the cortical region how best to represent stimuli. New hippocampal representations are then transferred incrementally to the internal layer nodes of cortical networks, one of which is shown in Figure 1A. The cortical network can then learn to map from these new hippocampal-trained representations to a behavioral response. In contrast to the hippocampal network, which requires a sophisticated multilayer learning algorithm, the cortical network may be trained by a simple correlation learning rule, such as the least mean squares (LMS) algorithm (Widrow and Hoff 1960).

Within this model framework, hippocampal-region damage is simulated by disabling the hippocampal-region network (Fig. 1B). In this case, the cortical network is unable to acquire new representations, although it can still learn new behavioral responses based on preexisting (and fixed) stimulus representations.

The cortico-hippocampal model of Figure 1 can be applied to simple associative learning paradigms, such as classical conditioning; it has proven to capture many significant trial-level conditioned behaviors in intact animals, and in animals with damage to the hippocampal region, including the hippocampal formation and overlying cortical input areas (Gluck and Myers 1993; Myers and Gluck 1994). We now review a few of the critical studies that support this theory.

**Evidence for Redundancy Compression**

The first putative hippocampal-dependent recoding (or representational) bias, redundancy compression, is a bias to increase similarity between co-occurring or redundant stimuli. One of the simplest paradigms in which redundancy compression is expected is sensory preconditioning. Consider two distinct stimulus cues, A and B, perhaps two tones; the representations they evoke should be highly distinct and, thus, there should be very little generalization between A and B. If A is subsequently paired with the US (A + training), a test presentation of B should evoke very little response. However, if prior to the A + training there are repeated nonreinforced trials pairing A and B, redundancy compression makes the representations of co-occurring cues A and B more similar. This will increase generalization between A and B, so that subsequent A + training will transfer partially to B, and thus a test presentation of B will evoke some conditioned responding. This sensory preconditioning effect is seen in intact animals (Thompson 1972) and also in the intact model (Fig. 2A). Because the Gluck and Myers model assumes that sensory preconditioning arises from hippocampal-dependent representational compression during the pretraining phase, the effect is not seen in the lesioned model (Fig. 2A). Similarly, hippocampal damage through fimbrial lesion attenuates sensory preconditioning in rabbits (Port and Patterson 1984).

Conversely, in compound preconditioning, preexposure to AB retards later learning to discriminate them in both normal animals and children (Lubow et al 1976). The intact model shows this behavior (Fig. 2B), because preexposure increases generalization and retards discrimination of A and B; this hippocampal-dependent effect
**Figure 2**: Simulation results with the intact and lesioned models. (A) Sensory preconditioning: Unreinforced preexposure to a compound AB followed by training to respond to A results in stronger responding to B alone than in a control condition with no preexposure. The intact but not lesioned model shows this effect; fimbrial lesion similarly eliminates sensory preconditioning in rabbits (Port and Patterson 1984). (Solid bars) A+ only; (shaded bars) AB-, then A+. (B) Compound preconditioning: Unreinforced preexposure to AB in the intact model slows later training to discriminate A and B, as shown by less relative difference in responding to A and B [Diff(A,B)]. Intact rats show compound preconditioning (Lubow et al. 1976); the model predicts hippocampal-region lesion should eliminate the effect. (Solid bars) A+, B- only; (shaded bars) AB-, then A+, B-.}

EVIDENCE FOR PREDICTIVE DIFFERENTIATION

In the Gluck and Myers theory, the hippocampal region is also assumed to mediate predictive differentiation, decreasing the similarity between stimuli that are to be mapped to different outputs. The simplest paradigm in which differentiation is expected is a discrimination task in which two stimuli, A and B, are associated with different responses (e.g., A predicts the US but B does not). In the intact model, the hippocampal region constructs new internal representations that decrease the similarity between A and B. These new differentiated representations are acquired by the cortical network’s hidden layer, which can then map these representations to different responses, as the task requires. Interestingly, the lesioned model shows no particular deficit on this task, and learns as quickly as the intact model (Fig. 3A; Gluck and Myers 1993). This is because the preexisting (fixed) hidden layer representations in the lesioned model are likely to partially distinguish A and B for this simple task, and thus all the network must do is map these distinguishable representations to the correct responses.

Consistent with the model’s implications, hippocampal lesions do not impair learning a simple discrimination across a variety of preparations (e.g., Silveira and Kimble 1968; Jones and Mishkin 1972; Ross et al. 1984; Port et al. 1986; Zola-Morgan and Squire 1986; Good and Honey 1991; Zola-Morgan et al. 1992; Jarrard 1993). In some cases, hippocampal-region damage has even been shown to facilitate learning (e.g., Schmaltz and Theios 1972; Port et al. 1985; Eichenbaum et al. 1986). Gluck and Myers’s cortico-hippocampal model gives an interpretation for this seemingly paradoxical facilitation of learning after hippocampal lesion. Figure 3A shows that the lesioned model learns somewhat faster than the intact model, which must spend time constructing new representations that are not strictly needed for the training task (but, as described later, these new representations may be relevant for future generalizations of this learning).

The additional time and effort spent by the intact model in constructing these new and differentiated stimulus representations can be very helpful if task demands change so that the same cues are relevant but their meaning changes. A simple example of this occurs in the easy/hard transfer paradigm, in which animals are first trained on an easy discrimination (e.g., black vs. white) and then transferred to a hard discrimination along the same stimulus continuum (e.g., dark gray vs. light gray). This transfer facilitates learning more than an equivalent amount of pretraining on the hard discrimination itself (e.g., Lawrence 1952; Terrace 1963; Riley 1968). The intact model correctly shows this effect (Fig. 3B; Gluck and Myers 1993): Diff(1+,1-) is a measure of the discrimination of the two hard stimuli, is greater following pretraining on the easy task than following pretraining on the hard task. During pretraining on the easy task, the hippocampal-region network is assumed to differentiate the representations of the two stimuli, which predict different outcomes. Because these two stimuli are assumed
A Discrimination  B Easy-hard transfer

Figure 3: Simulation results. (A) Stimulus discrimination: training to respond to A but not to B. The intact model forms new stimulus representations, which differentiate A and B, and then maps them to opposite responses. The lesioned model simply maps from preexisting (fixed) representations in the cortical network to the correct responses. As a result, there is no impairment in conditioned discrimination learning in the lesioned model. (B) Easy/hard transfer: Learning a hard discrimination between A and B is facilitated by prior training on an easier discrimination along the same stimulus continuum, in the intact but not lesioned model. This is shown by less relative difference in responding to A and B \[\text{Diff}(A,B)\] after a fixed amount of training in the lesioned model than in the intact model. The prediction of hippocampal dependence remains to be tested in animals. Reprinted, with permission, from Myers et al. (1995). (Solid bars) Easy/hard training; (shaded bars) hard/hard training.

This leads to the novel prediction that hippocampal-lesioned animals will not show easy/hard transfer, another prediction that remains to be tested.

The model also applies to data regarding the effects of hippocampal lesion on contextual processing. Many of the learning deficits associated with hippocampal damage can be described as contextual deficits, as they suggest an inability to incorporate information about the environmental conditions under which an event occurs (Hirsh 1974). It should be noted that this does not reflect a general inability to perceive contextual cues, because lesioned animals can still learn to discriminate contexts (e.g., Good and Honey 1991; Phillips and LeDoux 1994). What seems to be disrupted in the lesioned animal is the ability to use context to interpret the meaning of conditioned cues (Myers and Gluck 1994). Within the cortico-hippocampal model, the contextual information is included in the representations formed in the hippocampal region’s internal layer. As reviewed in Myers and Gluck (1994), the cortico-hippocampal model of Gluck and Myers (1993) can account for results from a range of context studies comparing intact and lesioned animals; moreover, it provides a computational instantiation and elaboration of several existing qualitative theories that have implicated the hippocampus in context learning (Hirsh 1974; Penick and Solomon 1991).

We turn now to Knowlton et al.’s (1994) data on category learning in amnesic patients, and show how the cortico-hippocampal model can account for these data as well.

PROBABILISTIC CATEGORY LEARNING IN AMNESIA

Knowlton et al. (1994) conducted three experiments in which amnesic patients and control subjects were given a series of multidimensional stimuli and asked to classify them into one of two categories. For example, in one task (Experiment 2) there were four cards, each with a unique set of abstract geometric patterns (see Fig. 4). The stimuli were drawings of up to four of these cards, and subjects were asked to predict whether there would be good or bad weather based on these cards. The actual weather was determined according to a probabilistic rule based on the cards, and thus each card was a partial predictor of the weather. Some cards were more accurate predictors than others. These probabilistic relationships between cues and outcomes ensured that it was
impossible for subjects to learn the categorization with complete certainty, although it was possible to use the probabilistic card–weather relationships to achieve significantly better-than-chance performance. The best possible performance in such a probabilistic category learning task is achieved by a subject who responds to each stimulus with the outcome most often associated with that stimulus.

Knowlton et al. (1994) evaluated their subjects' performance in terms of how frequently each subject made this optimal response. In each category learning task, the amnesic patients initially learned to associate the stimulus cues with the appropriate outcomes at the same rate as control subjects, improving from chance performance (50% correct) to ~65% correct over the first 50 trials. There are no significant differences between the normal and amnesics early during the first 50 trials, and the only significant differences were later in training. With extended training, however, control subjects eventually outperformed amnesic patients (see Fig. 5A). Knowlton et al. termed this an "early/late" effect, because the effects of hippocampal damage were apparent only late in training.

In a related earlier study, Squire and Frambach (1990) also found a similar late-training amnesic deficit in a similar probabilistic prediction paradigm involving a different cover story (Fig. 5B). Their subjects estimated the number of workers needed to keep sugar production at a constant level for a fictional factory; the rule relating sugar production and size of workforce was probabilistic. Squire and Frambach found that, whereas the amnesic patients' performance matched that of control subjects early in training, the control subjects surpassed the amnesic patients later in training. Thus, as in the Knowlton et al. study, control subjects and amnesic patients in the Squire and Frambach study show similar behavior early in training, but the controls outperform the amnesic patients later in training. As described in the next section, this late-training advantage for controls versus amnesics is of theoretical importance because it is at odds with the expectations of some prevailing interpretations of how independent memory systems might interact.

**EARLY VS. LATE DIFFERENCES IN THEORIES OF DECLARATIVE AND IMPLICIT MEMORY**

One popular interpretation of amnesia is that there are two independent memory systems: a “declarative” memory system and a “procedural or habit” memory system. Control subjects are expected to have both systems operating with one or the other (or both) determining behavior on any given memory task. Amnesic patients, however,
are expected to have access only to their procedural/habit memory system.

This view of memory as being decomposable into multiple independent processes might lead us to expect that early in training—when there are only a few trials to remember—we should find a facilitation for intact subjects who are able to use their declarative system to recall all or most of the previous training trials. Later in training, however, when the declarative system has stored far too many individual trials to accurately recall all of them, we might expect behavior to be controlled by the procedural/habit system. This expectation of an early-training advantage that might follow from some theories of independent and interacting dual memory processes is, however, just the opposite of what Knowlton et al. (1994) found in their study of amnesic category learning. As we show below, the Gluck and Myers (1993) theory of cortico-hippocampal function correctly expects that the advantage for intact subjects should occur later, rather than earlier, in incremental associative learning tasks like classical conditioning and probabilistic category learning.

APPLICATION OF CORTICO-HIPPOCAMPAL MODEL TO THE DATA OF KNOWLTON ET AL.

We now turn to show how the late-training deficit found in the Knowlton et al. (1994) study of amnesic category learning can be understood as a natural consequence of Gluck and Myers’s (1993) theory of hippocampal-region function, a theory that has heretofore been applied solely to studies of animal learning. Although the cortico-hippocampal model of Figure 1 was originally developed as a model of associative learning in animals, previous work in cognitive psychology suggests that elementary behavioral principles of conditioning can be generalized to account for a wide range of human category learning behaviors (Gluck and Bower 1988a,b). To apply the cortico-hippocampal model to the Knowlton et al. (1994) category learning task, the model was generalized to allow two output nodes, one corresponding to each possible classification output, and a simple probabilistic rule was used to generate the actual classificatory response on the basis of these network outputs (Fig. 6). Full details of the model used in these simulations are given in Materials and Methods.

When applied to the Knowlton et al. (1994) task, the cortico-hippocampal model produces learning curves analogous to those found in the human experiments. Figure 5C displays the learning curves produced by the lesioned and intact versions of the model. The results can be summarized as follows: early in training, the lesioned model’s performance matches that of the intact model ($F = 2.319$, $df = 1, 12, P = 0.154$), but later in training the performance of the intact model surpasses that of the lesioned model ($F = 23.248$, $df = 1, 12, P = 0.000$) although the performance of the lesioned model, like that of the amnesic patients, is above chance ($t = 5.225$, $d = 6, P = 0.002$ for the later part of training). This qualitative pattern of results in the model mimic that found by Knowlton et al.: no difference between control and amnesic patients during the early (first 50) trials, but a significant advantage for the controls over the amnesics late in training, with both groups performing significantly above chance.

Note that the fact that the lesioned model performs better than chance indicates that this probabilistic learning task can be learned, at least partially, on the basis of stimulus–response associations alone, without the influence of hippocampal-dependent recoding. The most important result from these simulations is that the late-training deficit in the lesioned model mimics the late-training deficit seen in the amnesic subjects.

As noted earlier, this late-training advantage for the intact model occurs because the hippocampal network develops new stimulus representations to aid learning; these representations can develop only over many trials as the model is exposed to a representative subset of training trials.
patterns. However, early in learning, before these representations develop, the intact model depends on the preexisting representations in the cortical network, and maps these to (an approximation of) the correct response. Later in training, as the hippocampal representations become available, these new representations are acquired by the cortical network and allow further improvement in performance. In the lesioned model, of course, only the prior cortical representations are ever available. Thus, early in training, we expect performance to be similar in both the lesioned and intact models; only later, as training continues, and no new representations become available to the lesioned cortical network, do we see the lesioned model showing a deficit relative to the intact model (Fig. 5C), just as seen in the data from Knowlton et al. (Fig. 5A). It is important to note that the main point of this simulation is to show how the model reproduces the late-training deficit found seen in the results of both Knowlton et al. (1994) and Squire and Frambach (1990); the model does not, however, provide a quantitative fit to the data.

FURTHER IMPLICATIONS OF PREDICTED EARLY VS. LATE DIFFERENCES FOR LEARNING AND GENERALIZATION

We now consider the implications of the cortico-hippocampal model’s early/late distinction in hippocampal function for another learning paradigm. In particular, the cortico-hippocampal model expects that the impact of hippocampal lesions on generalization from one task to another should depend on the amount of training experienced in the first task. As described below, this analysis suggests some possible interpretations for previously paradoxical and apparently contradictory results in the animal literature on stimulus generalization, and leads to further novel predictions of the theory.

STIMULUS GENERALIZATION

Studies of stimulus generalization provide an important source of data on the difference between early and late training. 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 a 1200 Hz tone with a subsequent shock US. The animal would then be tested with novel similar cues that vary along this continuum, such as tones of 400 Hz, 800 Hz, and 1600 Hz (Fig. 7A). 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 upward approximating a steep exponential-decay curve (Shepard 1958, 1987). This steep gradient is interpreted in the cortico-hippocampal theory as being a result, in part, of the stimulus recoding that enhances the difference between the training stimulus and all other

Figure 7: Stimulus generalization gradients produced by training a response to one cue, and then testing response to cues that progressively differ from the trained cue. (A) Normal rabbits (solid line) show a steep generalization curve, as evidenced by strong response to the conditioned stimulus (1200 Hz) that rapidly decreases for tones of differing frequency, whereas hippocampal-lesioned animals (dashed line) show a much broader generalization gradient, and continue to respond strongly to tones near the frequency of the conditioned stimulus (adapted and redrawn from Solomon and Moore (1975). (B) Similarly, the intact model (solid line) shows a steeper generalization gradient than the lesioned model (dashed line), reflecting the hippocampal-mediated process of differentiation. Reprinted, with permission, from Gluck and Myers (1993).
cues. This enhanced differentiation depresses generalization between the training stimulus and other stimuli, resulting in a steeper generalization gradient. Because this recoding is assumed to be hippocampal-dependent, the theory expects that hippocampal-lesioned animals should show broader generalization gradients than control animals (Gluck and Myers 1993). Indeed, Solomon and Moore (1975) showed that hippocampal-lesioned animals exhibit broader stimulus generalization gradients than intact control animals (Fig. 7A), just as expected by the model (Fig. 7B).

Our analyses suggest that this difference in generalization gradients should depend on the amount of training. Figure 8 shows generalization gradients produced by the intact and lesioned versions of the cortico-hippocampal model at various stages of training to the training stimulus. Generalization in the intact model (Fig. 8A) begins with a broad gradient reflecting the initial (prerecoding) state of the cortical representations. With subsequent training the hippocampal-dependent recoding enhances the differentiation between the training stimulus and all other stimuli, effectively sharpening the gradient into its final concave-upward shape, just as seen above in Figure 8. Thus, the cortico-hippocampal model expects a growth in the stimulus generalization gradients resulting from the interaction between two processes: a hippocampal-independent associative process and a hippocampal-dependent representational recoding process. This interaction causes the generalization curves to first broaden and then later sharpen, resulting in a nonmonotonic change in the gradient over time at various intermediate points along the curve. Note that this nonmonotonic growth over time of the generalization gradients in which the gradient is tuned, or sharpened, over training is exactly what Thompson (1958) found in his study of stimulus generalization at varying levels of training. This was, in the late 1950s, an apparently paradoxical result because the prevailing one-process theories of associative learning expected just the opposite result, namely, that generalization to a stimulus, $S'$, should increase monotonically over time with the amount of training given to a similar stimulus $S$. Thus, the cortico-hippocampal model successfully accounts for existing behavioral data on stimulus generalization and training level for normal intact animals.

The generalization gradients of the lesioned model with variable amounts of training are shown in Figure 8B. The lesioned model produces much broader gradients at all stages in training. As illustrated in Figure 7, the model predicts that asymptotically the gradients in the lesioned model should be broader, and less precisely tuned, than in the intact model. Moreover, the model makes the novel prediction that the observed differences in generalization gradients found by Solomon and Moore (1975) should emerge only late in training, once the hippocampal-dependent stimulus recoding has occurred. The relevant study—a variation on the Thompson (1958) experiment, comparing lesioned and control animals—remains to be done.
Materials and Methods

The stimulus pattern presented on each trial was represented by four binary-valued digits. When a stimulus cue was present the corresponding digit took the value 1, and when absent, the corresponding digit took the value 0. Category assignment was represented by two binary-valued digits. Category A was represented by 1 0; Category B by 0 1.

THE HIPPOCAMPAL MODEL

The hippocampal network consisted of six input nodes, three hidden nodes, and six output nodes. There was full connectivity between layers. Input to the hippocampal network consisted of a six-bit vector, comprised of the four stimulus digits and two zeroes corresponding to the unknown category value. Desired output consisted of a six-bit vector, comprised of the four stimulus digits and the two category digits. Node output was calculated as:

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

where \( f \) is the logistic function defined by

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

and \( w_{ij} \) is the weight from node \( i \) to node \( j \) and \( \theta_j \) is node \( j \)'s bias. Weights and biases were initialized according to a uniform distribution \( U(-3,3) \). Weight update occurred according to the error back-propagation algorithm (Rumelhart et al. 1986):

\[ \Delta w_{ij} = \beta \delta_j y_i + \alpha (\Delta w_{ij}) \]

where \( \delta_j = (l_j - y_j) y_j (1 - y_j) \) for output node \( j \) and

\[ \delta_j = y_j (1 - y_j) \left( \sum_k w_{jk} \delta_k \right) \]

for hidden node \( j \). \( l_j \) is the desired output for node \( j \). \( \beta \) was set at 1.0 throughout training; and \( \alpha \) was set to 0.9. \( \theta_j \) was trained as if it were a weight to node \( j \) from a node that constantly output 1.0.

THE CORTICAL MODEL

The cortical model consisted of four input nodes, 30 hidden nodes, and two output nodes. Input consisted of the four stimulus digits. Desired output was provided by the two category digits. Node output was calculated as in the hippocampal network. Weights and biases were initialized according to a uniform distribution \( U(-3,3) \). The weight from hidden layer node \( i \) to output node \( j \) was updated according to:

\[ \Delta w_{ij} = \beta (l_j - y_j) y_i \]

where \( \beta \) was set at 0.2 throughout training. There was full connection from the hidden layer to the output layer of the cortical model.

There was full connection from the hippocampal hidden layer to the cortical hidden layer, with weights initialized from \( U(-3,3) \). These weights were nonadaptive. The error signal to cortical hidden node \( j \) was calculated as:

\[ \delta_j = \sum_b y_b w_{bj} \]

where \( H \) is the number of hippocampal hidden units. The weights from cortical input node \( i \) to hidden layer node \( j \) were updated as

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

The learning rate \( \beta \) was set to 0.1 throughout training.

OUTPUT AVERAGING

We applied the intact and lesioned model to the seven stimulus orderings that Knowlton et al. (1994) presented to the amnesic subjects in their second task. For each ordering, 10 stimulations of each model were run. The activations for each of the two cortical output nodes were averaged over the 10 stimulations. From the average output activation for each ordering, the probability of a correct response was calculated. On any trial, a category was considered to be the correct response whenever that category was the most frequent response to the input pattern. The probability of a correct response, \( P(\text{Correct}) \), was calculated as the ratio of the activation correspond-
ing to the correct response, \( Act(Correct) \), over the sum of the output activations.

\[
P(Correct) = \frac{Act(Correct)}{Act(Correct) + Act(Incorrect)}
\]

DATA ANALYSIS

EARLY LEARNING

For the first 50 trials, the probability of a correct response produced by the lesioned and intact models did not differ \((F = 2.319, df = 1,12; P = 0.154)\). Performance was at the chance level for the first block of 10 trials \((t = 1.119, df = 13, P = 0.283)\), and above chance for the last block of 10 trials \((t = 5.080, df = 13, P = 0.000)\).

LATER LEARNING

When training was extended past 50 trials, differences between the intact and lesioned models emerged. The performance of the intact model surpassed that of the lesioned model during the latter part of training \((F = 23.248, df = 1,12; P = 0.000)\). The lesioned model continued to produce greater than chance performance \((t = 5.225, df = 6, P = 0.002)\).

Discussion

We have shown here that Gluck and Myers’s (1993) cortico-hippocampal model expects a late-training deficit for amnesic patients in multiple-cue probabilistic category learning, just as was found by Knowlton et al. (1994). We attribute the late training deficit in amnesic patients to a hippocampal-dependent recoding of stimulus representations that requires a sufficient number of training trials before regularities in stimulus–stimulus relationships become apparent. These representations are then passed on to the cortical structures responsible for the formation of stimulus–response associations. Without access to the hippocampal-dependent stimulus representations, learning still proceeds, but based only on a fixed set of representations. In further explorations of this early versus late distinction, we have shown how this analysis of hippocampal function might lead to insights into the effects of training level on analyses of stimulus generalization. Although the data that do exist on this issue within the animal literature are consistent with our analyses, a considerable number of further experiments will be required to evaluate these predictions more completely.

Our application of Gluck and Myers’s (1993) cortico-hippocampal model to Knowlton et al.’s (1994) study of category learning in amnesia represents only a preliminary attempt to extend the model to human category learning. Further analyses of the model will be required to assess whether or not it can account for the broad spectrum of existing data on human category learning. Nevertheless, our analyses imply that careful comparison between early- and late-training differences in learning may be important for understanding amnesia and hippocampal function.

More generally, we see these results as supporting the need for a closer rapprochement between the fundamental interests and behavioral paradigms of neuropsychological studies of memory, and the methods, models, and behavioral paradigms of cognitive psychology. For example, cognitive neuroscience approaches to memory research might benefit from the quantitative behavioral analyses and formal computational models that have emerged in cognitive psychology from detailed parametric studies of category learning.

Similarly, researchers studying the cognitive psychology of human learning might benefit from an infusion of data, ideas, and constraints from cognitive neuroscience. In recent years cognitive psychology has seen a profusion of learning models that are all quite successful at accounting for a wide range of human learning behavior, although they may adopt vastly different frameworks and processing assumptions (e.g., Medin and Schaffer 1978; Estes 1986; Gluck and Bower 1988b; Nosofsky 1989; Kruschke 1992). The concurrent existence of numerous successful models might be interpreted as suggesting that the field has run out of important data to explain, forcing researchers to focus on increasingly finer grained aspects of quantitative details. It is quite possible, however, that further behavioral analyses to discriminate among these models may be futile given the inherent limitations of the behavioral data alone. Perhaps only with the introduction of new data and constraints from neuropsychological studies of memory-impaired patients will researchers be able to distinguish between otherwise equally adequate behavioral models.
Acknowledgments

For their helpful comments and suggestions on early drafts of this manuscript, we are grateful to Brandon Ermita, Valerie Hutchison, Barbara Knowlton, John Moore, Dawn Radice, and Larry Squire. This research was supported by grants to M.A.G. from the Office of Naval Research through the Young Investigator program and by grant N00014-88-K-0112, as well as a grant from the McDonnell-Pew Program in Cognitive Neuroscience.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked ‘advertisement’ in accordance with 18 USC section 1734 solely to indicate this fact.

References


Thompson, R. 1958. Primary stimulus generalization as a...


Received November 8, 1995; accepted in revised form October 4, 1996.
Late-training amnesic deficits in probabilistic category learning: a neurocomputational analysis.

M A Gluck, L M Oliver and C E Myers

Access the most recent version at doi:10.1101/Im.3.4.326

References

This article cites 53 articles, 6 of which can be accessed free at:
http://learnmem.cshlp.org/content/3/4/326.full.html#ref-list-1

Email Alerting Service

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or click here.