HIPPOCAMPUS 13:299–312 (2003)

Transitivity, Flexibility, Conjunctive Representations, and the Hippocampus. II. A Computational Analysis

Michael J. Frank, Jerry W. Rudy, and
Randall C. O'Reilly*

Department of Psychology, University of Colorado, Boulder, Colorado

ABSTRACT: A computational neural network model is presented that explains how the hippocampus can contribute to transitive inference performance observed in rats (Dusek and Eichenbaum, 1997. Proc Natl Acad Sci U S A 94:7109–7114; Van Elzakker et al., 2003. Hippocampus 12:this issue). In contrast to existing theories that emphasize the idea that the hippocampus contributes by flexibly relating previously encoded memories, we find that the hippocampus contributes by altering the elemental associative weights of individual stimulus elements during learning. We use this model to account for a range of existing data and to make a number of distinctive predictions that clearly contrast these two views. *Hippocampus 2003;13:299–312.* © 2003 Wiley-Liss, Inc.

KEY WORDS: transitive inference; hippocampus; conjunctive representations; computational models; learning

INTRODUCTION

Rats, pigeons, and primates display what is referred to as transitive inference. In everyday terms, transitive inference is used when one is told that John is taller than Bill, who is taller than Fred, and then one logically infers that John is also taller than Fred. As studied in the laboratory, this phenomenon emerges after subjects have been trained on a series of simultaneous discrimination problems involving common stimuli, e.g., A+B−, B+C−, C+D−, D+E−, where + designates the rewarded choice, and − designates the nonrewarded choice. After such training, the subject is then tested with a novel combination, BD, and transitive inference is demonstrated when the subject can "infer" that because B is chosen over C, and C is chosen over D, then B must be chosen over D. This outcome has been obtained a number times (e.g., von Fersen et al., 1991; Dusek and Eichenbaum, 1997; Van Elzakker et al., 2003) in neurologically intact subjects.

The central problem addressed in this article is the fact that rats with hippocampal damage can solve the training problems but fail the BD transfer test (Dusek and Eichenbaum, 1997). Dusek and Eichenbaum (1997) interpreted their results as strong support for the idea that the hippocampus provides the substrate for representational flexibility—that information stored in the hippocampus can be retrieved and used appropriately in novel situations. Specifically, they proposed that the intact rat encoded a relational representation of the individual elements of problems, A>B>C>D>E, that enabled comparisons to be made among them. Thus, if confronted with the novel combination, BD, the rat compares the position of B and D on the ordered representation and infers that if B>C and C>D, then B>D. This comparison leads to a choice of B.

We noted in our companion article (Van Elzakker et al., 2003), however, that there is good reason to question the representational flexibility account of transitive inference based on results from rats trained on a five-problem discrimination set (A+B−, B+C−, C+D−, D+E−, E+F−). This five-premise version of the problem permits two tests of transitivity, BD and BE. Our rats, however, only displayed transitivity when tested with BE. In contrast, the representational flexibility account would appear to predict that rats should display transitivity when tested with either novel pair. Our analysis of these results led us to conclude that transitivity was not mediated by the inferential-like processes needed for representational flexibility, but instead was the result of rats making their choice based on the absolute associative excitatory strength of the individual test cues. The concept of associative strength represents the capacity of the stimulus to evoke a choice response.

In essence, our analysis implies that the previous conclusion that rats are using inferential-like processes to make their choice resulted because the novel test did not satisfy the requirement that the choice cues have equal associative strength and therefore did not provide a true test of transitive inference. Our account shares common ground with other theorists who have addressed this problem (e.g., von Fersen et al., 1991; Siemann and Delius, 1998).

Our analysis and conclusions raise two issues for a theory of transitivity. First, why does training on a set of premise pairs (A+B−, B+C−, C+D−, D+E−, E+F−) result in a graded level of associative strength among individual stimuli that on the surface should have

Grant sponsor: National Institute of Mental Health; Grant number: MH613616.
*Correspondence to: Randall O'Reilly, Department of Psychology, University of Colorado, Boulder, 345 UCB, Boulder, CO 80309.
E-mail: oreilly@psych.colorado.edu
Accepted for publication 22 May 2002
DOI 10.1002/hipo.10084

equal values (e.g., B, D, and E)? Second, how does the hippocampus contribute to this outcome? The purpose of this article is to address these two questions. To do this, we use a biologically based model of the hippocampus and neocortex that we have previously applied to a variety of phenomena (O'Reilly and Rudy, 2001).

This article proceeds in several stages. First we establish our position that transitivity is a product of graded associative strengths to the individual cues that develop during acquisition, and not a product of representation flexibility. We then describe in detail how the computational model is implemented. We then show (1) that this model produces graded associative strengths to the individual cues that can mediate transitivity, and (2) how the hippocampus contributes to graded associative strength. We also derive several testable predictions from the model and conclude with a discussion on the relationship of our model to other views.

GRADIENT OF ASSOCIATIVE STRENGTH

The representational flexibility hypothesis assumes that transitive behavior in nonverbal subjects is the product of a logical inference of the sort: if A>B and B>B, then A>C. Consequently, a valid test of this hypothesis requires that the reward associations of the individual stimuli (e.g., B and D) must be equal. Otherwise, subjects could be performing on the basis of these unequal associative values, instead of relying on relational flexibility. This is precisely the reason that the edge stimuli, A and E, are not used in transitive tests—A is always rewarded and E is never rewarded, so it would be trivial for the rat to choose A over E.

The results of the companion article (Van Elzakker et al., 2003) and the previous findings of von Fersen et al. (1991) with pigeons, however, suggest that the assumption of equal associative strength for the internal stimuli (including B and D) is false. Instead, these findings suggest a simpler explanation: The discrimination training procedures that establish the basis for performance on the test trials produces what we call a gradient of associative strength (consistent with similar ideas proposed by von Fersen et al., 1991; Siemann and Delius, 1998). This means that training produces unequal values of reward associations for the individual stimuli. It is this variation in the associative values of the stimuli, and not logical reasoning, that then dictates choice behavior on the transitivity test. Specifically, we show that the strong reward association of A (which is always rewarded) can effectively bleed over into B (as explained in more mechanistic terms in the next section), such that B also has a somewhat positive association. Similarly, the negative association with E (which is never rewarded) can bleed over into D, such that it also has a negative association. Therefore, when given the BD probe, the subject need not rely on comparative processes any more than in the AE case, as the individual associative strengths are sufficient to determine the selection of B over D. We refer to this bleeding-over from the "anchor" stimuli of A and E as the anchoring effect.

The empirical support for the idea of a gradient of associative strength emerged in the companion article, when rats were trained with five discrimination pairs (A+B−, B+C−, C+D−, D+E−, E+F−). This training permitted two tests of transitivity between two sets of internal stimulus pairs (BD and BE). The representational flexibility account predicts that subjects should choose B over both D and E. However, they did not: B was chosen over E, but not over D. Although these results are difficult to reconcile with the representational flexibility account, they are very consistent with the gradient of associative strength view (see also von Fersen et al., 1991). Simply put, B's associative value is greater than E's, but is similar to D's. Specifically, we suggest that the negative association from F bleeds over into E, but not sufficiently into D.

Thus, our computational model must accomplish two goals: (1) it must explain how training produces a gradient of associative values that is consistent with the data, and (2) it must clarify the contribution the hippocampus makes to transitivity. The representational flexibility account assigns its contribution to retrieval processes operating during testing. Our account will show that when the hippocampus contributes to transitivity, it does so during the acquisition phase when the associative values of the individual elements of the problem are being learned.

IMPLEMENTATION OF TRANSITIVE INFERENCE PARADIGM IN THE MODEL

We begin our explanation of the model by providing a high-level description of the kinds of processes that enable the rat to solve the premise discrimination problems. We then formalize these processes in our computational neural network model of the hippocampus and neocortex (O'Reilly and Rudy, 2001).

Experimental Paradigm

In the actual experiments (Dusek and Eichenbaum, 1997; Van Elzakker et al., 2003), the odors were presented to the rats in two sand-filled cups that were close together. A reward (+, Froot Loop) was buried in one of the cups and the rat could obtain it by digging it out of the sand. Thus, when the rat approached a cup, it had two options: it could dig or switch to the other cup. We used the five-problem (premise) set (A+B−, B+C−, C+D−, D+E−, E+F−) to simulate the problem because it permitted more than one test of transitivity: BD, BE, and CE.

We assumed that the choice behavior displayed by the rat is determined by two learning processes: (1) a stimulus selection process, and (2) a response selection process (Fig. 1). For example, when the rat encounters the first anchor choice stimuli (AB), it will sample a roughly equal combination of both odor stimuli and make an initial assessment of which stimulus to approach. Presumably this assessment is based on which stimulus is more likely to provide reward. This is the stimulus selection process. If the rat then approaches A, it will receive increased stimulation from A as compared with stimulus B, which we denote as Ab (where the closer stimulus is in upper case, and the further one in lower case). Conversely, if it approaches B, it would experience the Ba situation. The response selection process then operates on this config-

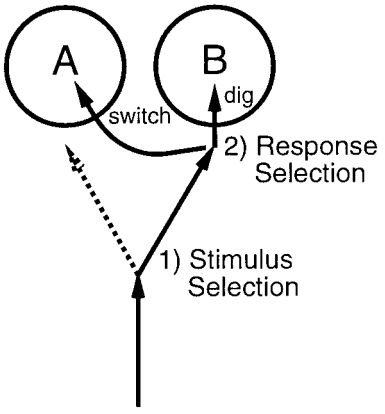

FIGURE 1. Two stages of selection as represented in the model. In the first stimulus selection stage, one of the two stimuli is selected to approach. In the second response selection stage, the decision to either dig at the selected (close) stimulus or switch to the other (far) stimulus is made. A and B represent two odors.

uration of the inputs: the rat decides either to dig for the reward at the closer stimulus or to switch to the other stimulus (in which case the rat digs in this other stimulus cup). Thus, in the Ab case, the rat could either dig at A or switch to dig at b, and the reinforcement outcome would be a direct consequence of this choice (i.e., it would be rewarded for digging at A, and not for switching to b). Note that this switching behavior was frequently observed during training—rats would approach one cup but then not dig, choosing instead to dig in the other cup.

These two processes are differentially engaged by the five training pairs (A+B−, B+C−, C+D−, D+E−, E+F−), which present qualitatively different problems to the rat. The anchor problems (A+B− and E+F−) can be solved just based on the differential reinforcement histories of the stimuli, and thus load more on the stimulus selection process. In the A+B− case, A is always rewarded, but B is rewarded only on half the trials, so the rat can relatively easily choose to approach A in the stimulus selection process. Similarly, in the E+F− case E is rewarded on one-half of the trials, but F is never rewarded, so the rat can solve this problem just by learning to avoid F in the stimulus selection phase. These anchor problems can be contrasted with the internal problems (B+C−, C+D−, D+E−). Because the rat is rewarded equally often for each of the individual stimuli in these problems, the rat would seem to have no straightforward solution. Thus, the stimulus selection process does not provide the solution to these problems, leaving it to the response selection process to make the final decision of where to dig.

In summary, the anchor and internal problems differ qualitatively because the anchor problems permit an elemental solution, and thus can be solved simply on the basis of the associative values attached to the individual cues. In contrast, the internal problems cannot be solved unless the subject constructs a conjunctive representation of the choice stimuli that resolves the ambiguity of their associative values (Rudy and Sutherland, 1995; O'Reilly and Rudy, 2001). In short, the stimulus selection process acts as an optimization mechanism (i.e., leading to faster performance) for the relatively easy anchor problems, but not for the internal problems. This fact will play a critical role in our simulations.

Implementation of Stimulus Selection

We separated the stimulus selection and response selection processes by treating stimulus selection as a mechanism that controls the probabilities of input patterns that are then presented to our standard neural network model of the neocortex and hippocampus (O'Reilly and Rudy, 2001). Thus, the neural network model constitutes the response selection mechanism (as described in the next section), and the stimulus selection mechanism simply controls the frequencies of input patterns that are presented to the network. For example, the stimulus selection process can turn the A+B− problem into two different response selection problems, Ab (A closer than b) or Ba (B closer than a) (Fig. 2). Consistent with our analysis of the behavior of the rat, we based the probabilities of stimulus selection for each of these patterns on the relative reinforcement values of the two choice stimuli:

$$\begin{aligned} P(Ab) &= Rew(A)/[Rew(A) + Rew(B)] \\ P(Ba) &= Rew(B)/[Rew(A) + Rew(B)] \end{aligned} \quad (1)$$

Thus, the probability with which the stimulus selection process selected Ab for the A+B− problem (P(Ab)) is simply the relative reward strength of A [Rew(A)] normalized by the total reward strength of both A and B.

Initially, the reward strengths of all the stimuli are set to 0.5, and therefore the probabilities are also 0.5. As the simulated rat experiences different patterns of reward for selecting the different stimuli, the reward strengths are modified, and this changes the probabilities of the two alternative presentations of a problem. Specifically, when any given trial is rewarded, the reward values of both the proximal and distal stimuli are modified. For example, if the stimulus selection process chooses to approach the A stimulus on the A+B− problem (Ab), and the response selection process chooses to dig in response to Ab, then a reward will be obtained. This reward will increment the reward strength of A [Rew(A)] by a small fixed amount (0.035), and decrement the reward strength of B by this same amount (i.e., the approach A and avoid B decisions are both rewarded, because both lead to the same outcome and there is no basis to favor one over the other). This will then increase the probability of Ab being presented next time (and decrease the probability of Ba) according to equation 0. Similarly, if the stimulus selection process selects Ba and the response process decides to switch instead of dig, then this is also rewarded and the reward values are incremented as before. If no reward occurs (because of either an incorrect dig or switch) then we did not change the values. We also tried an alternative scheme where the values were adapted in the opposite direction of the rewarded case, but with one-half the magnitude, and it made no significant difference to the results, so we used the simpler reward-only model. In sum, the change in reward strength for an item X is

$$\Delta Rew(X) = \begin{cases} +\ \lambda \text{ if } X \text{ selected and reward occurs} \\ -\ \lambda \text{ if } X \text{ not selected and reward occurs} \end{cases} \quad (2)$$

To see why stimulus selection is an important part of the model, we extrapolate to what will happen in the anchor cases of A+B− and E+F−. Because A is always rewarded, its reward strength Rew(A) will approach 1. However, Rew(B) will remain around 0.5 because

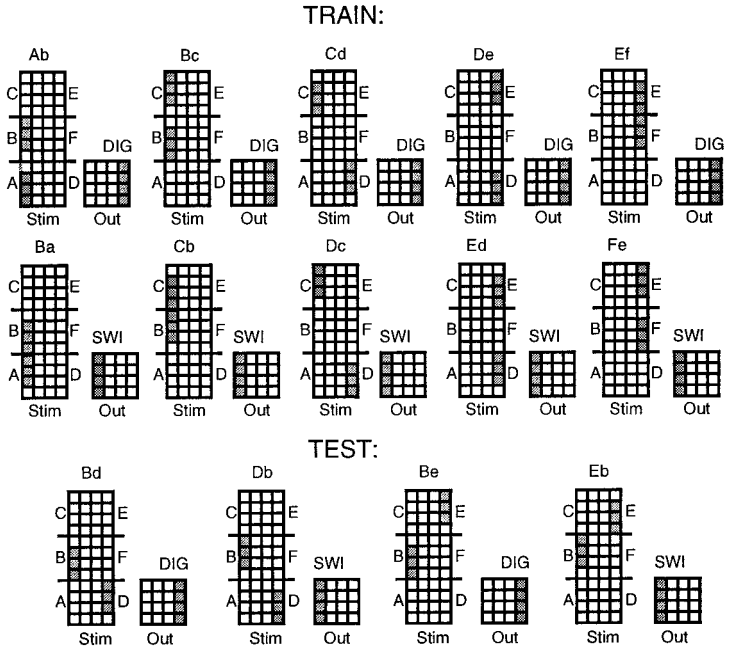

FIGURE 2. Training patterns: stimulus input pairs and trained output response. Each stimulus has four associated units, with the lower three units representing the proximal version of the stimulus (the stimulus that is initially approached by the stimulus selection process), and the upper three units representing the distal version. The model has to respond with Dig or Switch, with respect to the proximal stimulus. Note that the two versions of each stimulus overlap, with only one bit of information distinguishing the close and far representations.

it is equally often rewarded as not. The overall result is that the stimulus selection process will tend to approach A (i.e., produce Ab) two-thirds of the time (by equation 1: P(Ab) = 1/(1+0.5) = 0.67). Conversely, the selection of Ba will occur only one-third of the time. The net result is that the response selection model learns mostly to dig at the Ab pattern and does not have to learn as much to switch when faced with Ba. Consequently, it is easier for the model to learn to dig at the Bc pattern in the B+C− problem, because it does not have to overcome such a strong bias to switch in response to the B stimulus. Thus, stimulus selection provides a means for the "bleed-over" of the anchoring effects to neighboring stimuli. These effects are even stronger in the E+F− case, because F achieves a reward strength of 0, while E stays at around 0.5. Using equation 1, this means that after enough training Ef will come to be selected virtually 100% of the time (0.5/(0.5+0)), and Fe 0% [0/(0.5+0)]. This amounts to the rat always avoiding F, which is more extreme than the Ab case, in which the rat still approaches B one-third of the time. Because the elemental components of the interior problems (B+C−, C+D−,and D+E−) are all associated equally often with reward and nonreward, no dominant stimulus selection pattern will emerge, and both patterns associated with each problem will be presented about equally often to the response selection network.

Although the stimulus selection process is critical to the acquisition phase of the simulation, it is not implemented during the test phase when the model is presented with BD and BE. Thus, for example, when the rat is presented with BE, test performance is determined by the tendency of the network to produce a dig response when it is presented with the Be pattern and to generate a switch response when presented Eb. This is simply because the response selection model has no stochastic component and can be assessed with a single test, whereas the stimulus selection process is stochastic and would require sampling to minimize added noise in the measurements. Furthermore, because the tested elements typically have roughly equal associative value, the stimulus selection process would not substantially affect the results in any case.

Implementation of Response Selection

The response selection process is simulated via a neural network model (Fig. 3) that interfaces to the stimulus selection process by receiving the selected input patterns, and producing either dig or switch output patterns (Fig. 2). This model is our "standard model" of the hippocampus and neocortex (O'Reilly et al., 1998; O'Reilly and Munakata, 2000; O'Reilly and Rudy, 2001; Rudy and O'Reilly, 2001) The basic mechanisms of this model incorporate widely accepted ideas of cortical function in one coherent framework called Leabra (O'Reilly, 1998; O'Reilly and Munakata, 2000). The hippocampal component uses these same mechanisms, while incorporating the specialized anatomy and physiology of the hippocampus as emphasized by a number of theorists (Marr, 1971; McNaughton and Morris, 1987; Rolls, 1989; O'Reilly and McClelland, 1994; McClelland et al., 1995; Hasselmo, 1995).

The model learns to associate input patterns with output patterns. This learning is governed by a combination of Hebbian and error-driven learning, a central feature of the Leabra algorithm. The error-driven learning uses bidirectional activation propagation to communicate error signals, and is compatible with known

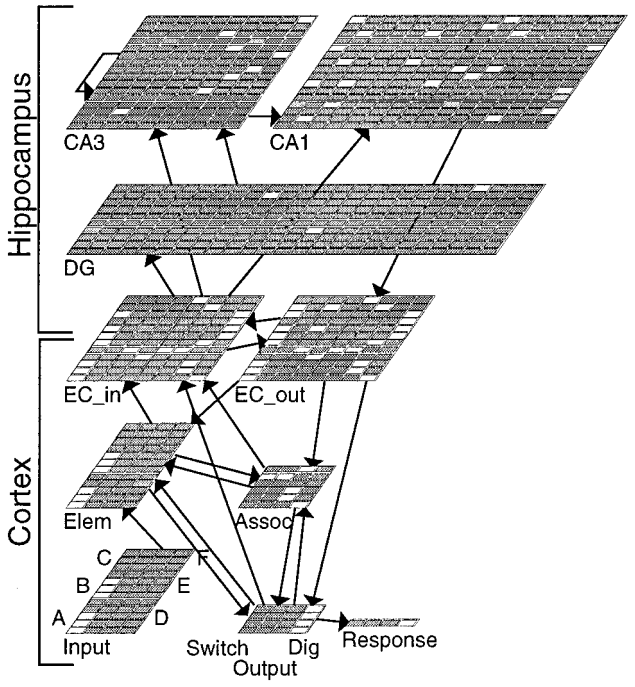

FIGURE 3. Neural network model of the response selection process, having cortex and hippocampus components. The activation pattern shown (active units in white) shows the Ab input pattern, and the response of the model is to Dig.

properties of synaptic modification (O'Reilly, 1996; O'Reilly and Munakata, 2000). The cortex has both an elemental representation of stimulus inputs (layer Elem in Fig. 3) and an associative layer that can learn higher-order representations of stimulus elements (layer Assoc in Fig. 3). Error-driven learning can, in response to explicit task demands over many trials of learning, shape representations in the associative layer to encode conjunctive representations of stimuli (O'Reilly and Rudy, 2001), as are needed to learn the internal problems. The output layer receives from both elemental and associative layers, and has two valid patterns: activation of all the units at the right side of the layer means "dig," while the left side means "switch." The response layer provides a simple reduction of the distributed output layer for coding responses: the output layer itself has to be distributed to provide a substantial representation in the hippocampal system.

The hippocampal system receives inputs (via entorhinal cortex (EC) layer EC_in) from the entire cortical system, including the output layer, and sends reciprocal projections back out to these same layers (via layer EC_out). The areas of the hippocampus proper (dentate gyrus (DG), area CA3, and area CA1) form conjunctive encodings over the EC inputs, and can contribute to output responses by associating a pattern of input cues with a pattern of activation over the output layer. The sparse activations of the DG, CA3, and CA1 layers produce a natural tendency for this system to encode stimulus conjunctions (Marr, 1971; O'Reilly and McClelland, 1994; O'Reilly and Rudy, 2001). The recurrent collaterals of area CA3 support pattern completion where a partial input pattern (e.g., the Ab input pattern) can trigger recall of the entire associated pattern that was previously conjunctively encoded (e.g., the "Dig" output pattern) (McNaughton and Morris, 1987; Rolls, 1989).

The model was initially trained by presenting the problems in blocks of trials: a block of A+B−, followed by a block of B+C−, and so on through E+F−. After the model was performing well in the blocked phase, the problems were randomly interleaved. This training sequence captures the essence of the training paradigm used in rats (Dusek and Eichenbaum, 1998; Van Elzakker et al., 2003). Our basic simulations used two blocks of seven trials per block, followed by 10 epochs (passes through all training cases) in the interleaved condition. After training, the model was tested with new combinations of stimuli (BD, BE, and AF). The number of trials per block, the total number of blocks, and the number of interleaved trials were all varied in order to analyze their effects on performance. Both the complete model and an incomplete model that lacked the hippocampal circuitry were trained.

It is critical to appreciate that the performance of the model, i.e., its output representation (Fig. 3), is determined by the combined influence of three systems: (1) elemental cortex, (3) associative cortex, and (3) the hippocampus. The results of our model can be understood as a product of the interaction of this feature of the architecture and error driven learning operating throughout the network. Note that the response layer is the source of the error signals that drive this learning. If there is no error on a given trial, then this form of learning is not present. Because the output of the model is determined by the summation of the three systems, a correct output that is produced by any one of the systems will eliminate error signals experienced by all systems. In other words, the network can exhibit a blocking effect like that described by the well-known Rescorla and Wagner (1972) model of Pavlovian conditioning. This model assumes that increments in associative strength are a product of the difference between the total associative strength of all cues present on a training trial and the asymptotic level of associative strength the unconditioned stimulus (US) will support. The blocking effect occurs when prior training to one stimulus element (A–US) prevents conditioning to another stimulus element (B) when the two stimuli are paired together with the US (AB–US) (Kamin, 1968). Because A already predicts the US, the error in prediction that drives new increments in associative strength is eliminated. Consequently, there is little or no increment in associative strength to the added cue, B.

A major goal of the simulations is to understand how the hippocampus contributes to transitivity. To preview the results, the simulations indicate that there is a "blocking effect" produced by output of the hippocampus that reduces what is learned by the other systems in the model.

Simulation of the Basic Transitivity Findings

We begin by describing how the intact and hippocampally lesioned models perform on the basic transitive inference task. Figure 4 displays the performance of the intact model during the final stage of interleaved training before the model was tested with the BD, BE, and AF probe stimuli. The lesioned model performance was indistinguishable from the intact model. Note that although the internal training problems require conjunctive representations,

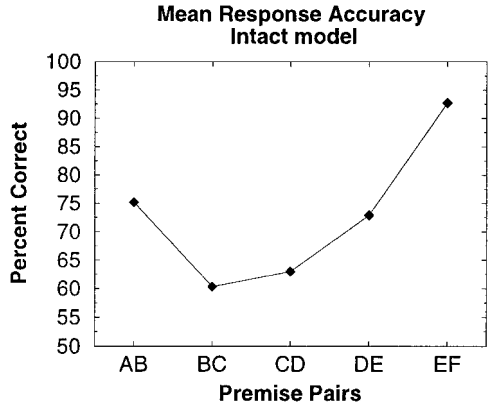

FIGURE 4. Errors during interleaved training. Note that the performance curve has the same shape as that of the behavioral rats in Eichenbaum's study. Specifically, performance during interleaved training is best for the E vs F pair; this has the least conflict because of the larger anchoring effect.

the lesioned model solved them. This outcome has been demonstrated in earlier work with this model (O'Reilly and Rudy, 2001). Furthermore, both models showed a strong anchoring effect, with performance being better for the anchor problems of A+B− and E+F− (which was nearly perfect). Note that the asymmetry in the AB versus EF problems is in accord with the stimulus selection asymmetry described previously, because B has some reward value to compete with A, but F has zero reward value and thus provides no competition with E. These results are similar to those found in rats and pigeons (Dusek and Eichenbaum, 1997; von Fersen et al., 1991; Van Elzakker et al., 2003).

Figure 5 displays the performance of the intact and lesioned models on three transfer problems: BD, BE, and AF. Figure 5 demonstrates two important phenomena:

1. Neither the intact nor lesioned model demonstrates transitive inference on the BD problem—they choose B over D only barely over the 50% chance level. This simulates the findings of Van Elzakker et al. (2003) and clearly shows that one cannot account for performance on this task in terms of logical inference.
2. The intact model displayed transitivity on the BE problem, reliably choosing B. In contrast, the lesioned network performed near-chance on the BE problem. This outcome simulates Dusek and Eichenbaum's (1997) finding that rats with damage to the hippocampal formation perform at chance on the transitivity test. It should be appreciated that Dusek and Eichenbaum trained rats on a four-problem set (A+B−, B+C−, C+D−, D+E−) and tested them on BD. However, just as E is a member of the anchor E+F problem in the five-problem set used to train the model, D in is a member of the D+E− anchor problem in the four-problem set used by Dusek and Eichenbaum. We will show that it is how the anchor problem is learned that is critical for understanding the contribution of the hippocampus; therefore, under our model, Dusek and Eichenbaum's BD is functionally equivalent to our BE.

Figure 5 also shows that both the intact and lesioned model reliably chose A on the AF trial. This outcome correctly simulates the results reported by Van Elzakker et al. (2003) and Dusek and Eichenbaum (1997). This is the expected outcome because of the different reinforcement histories associated with A and F; responding to A was always rewarded, but responding to F was never rewarded.

Analysis of Initial Results

On the test trials, the model is presented with novel combinations of stimuli. Consequently, conjunctive representations of the test pairs are not available, and we assume that the model must use the output of the elemental associative system to generate its choice. Thus, performance on the AF, BD, and BE test cases must be due to the individual elements' association with the dig or switch response.

Associative Weights

To examine the basis of responding to the other test cases, we examined the connection strengths between the elemental stimu-

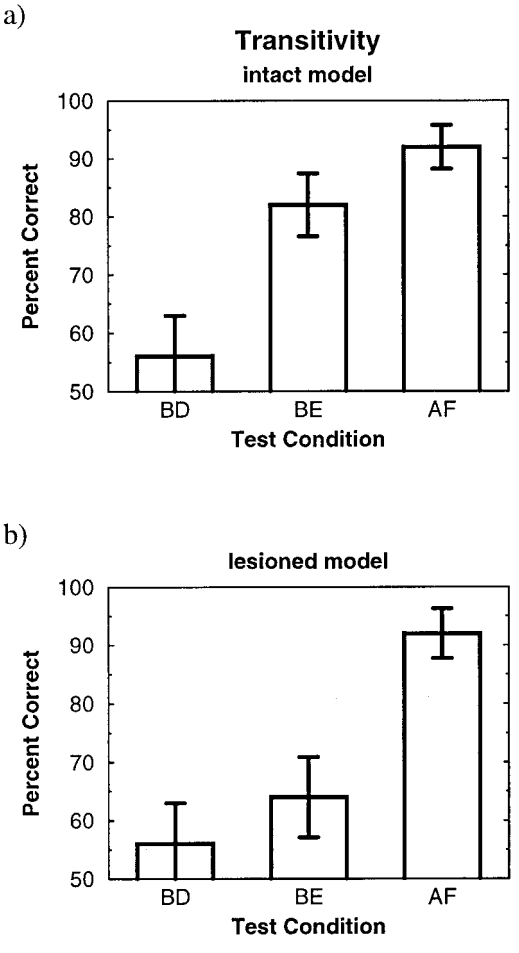

FIGURE 5. Transitive inference transfer pair performance for intact (a) and lesioned (b) models. Both models perform near-chance on BD, and near-perfect on AF (which is supported by strong elemental associations, given that A was always rewarded and F never was). The key finding is that intact model performs much better on the BE transitivity test than does the lesioned model.

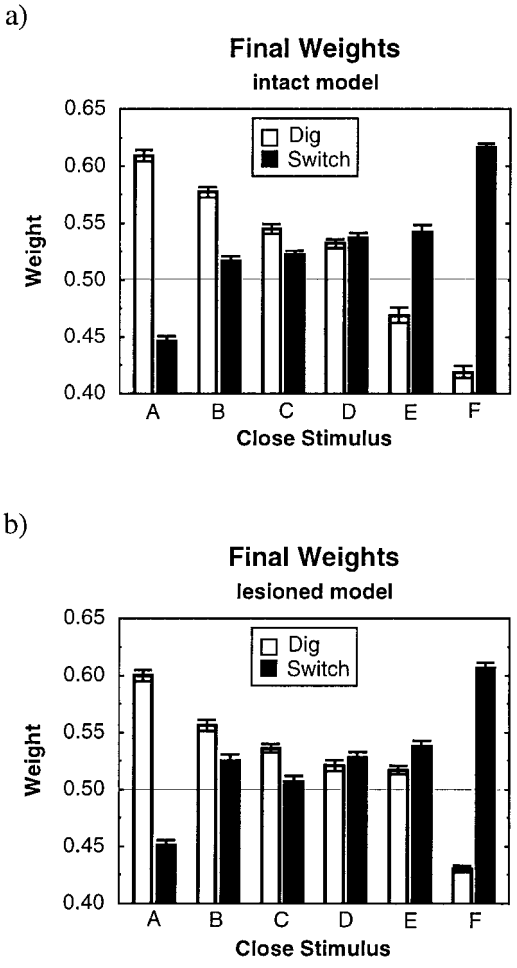

FIGURE 6. Averaged final weights for digging and switching in proximal stimuli after training for intact (a) and lesioned (b) models. The poor performance on BD by both models can be explained by noting that the model likes to both dig and switch to the D stimulus, producing random behavior even though it has a slight preference for digging in B. The differential performance on the BE problem arises because the intact model learns a net switch association from E (i.e., switch is significantly stronger than dig), which favors choosing B over E. Conversely, the lesioned model learns roughly equally strong dig and switch associations for E. Similarly, the intact model develops a stronger dig association for B than the lesioned model, which also produces better BE test performance in the intact model.

lus representations and the output responses of dig or switch (Fig. 6). By looking at the difference in strength between the dig versus switch response for the representation of the stimulus when it has been selected by the stimulus selection process (i.e., when it is the close stimulus), we can assess how the model will perform on test cases involving these stimuli. Complementary patterns of weights were found for the far representation of stimuli (i.e., close A is strongly associated with dig, while far a is strongly associated with switch).

We can directly explain the two key findings listed above in terms of these weight patterns:

Poor BD performance: Both the intact and lesioned model have ambiguous weights for the D stimulus—it is not strongly associated with either dig or switch responses, and thus could produce either response. Although the B stimulus is more strongly associated with digging, this is not enough to overcome the variable responses to the D stimulus. Thus, neither network exhibits a reliable overall pattern of "transitive inference" performance on the BD test case.

Differential BE performance: The intact network develops a stronger switch versus dig weights from the E stimulus. This means that it will tend to avoid E in the BE test, and as observed choose B over E. In contrast, the lesioned network has a much more ambiguous association with dig and switch, with only a weak overall switch association. Thus, it will perform poorly on the BE test. Furthermore, the intact model produces a stronger B dig association than the lesioned model, which also favors the transitive inference choice of B over E. Thus, the key to understanding the contribution of the hippocampal system on this problem is to understand how it contributes to the stronger E-switch and B-dig associations.

Hippocampal Blocking Effect

We noted that the blocking effect is central to understanding how the hippocampus contributes to the ability of the rat to display transitivity. From our analysis of the weights, we can conclude that this blocking effect is taking place primarily on the E+F− trials when the network correctly chooses to dig in E. Instead of this pattern producing a strong positive association between E and dig in the intact network, this association is blocked. The lesioned model does not have this blocking effect, and it therefore develops the stronger E-dig association that prevents it from choosing B over E on the BE test case. Furthermore, the stimulus selection process ensures that the network will tend to approach E much more frequently than F (because F is never rewarded), so there should be many opportunities for the E dig association to strengthen.

Why does the hippocampus block this E-dig association? First, unlike the cortex, the hippocampus quickly and automatically constructs conjunctive representations of co-occurring stimulus patterns (O'Reilly and Rudy, 2001). So, even though the solution to the E+F− anchor problem does not require a conjunctive representation, the hippocampal system of the intact model generates one automatically. For example, when the stimulus selection process presents the network with the close E, far F (Ef) pattern, the intact hippocampus constructs a conjunctive representation of that pattern and associates it with the dig response. Thus, because this hippocampal representation can produce the correct response on these trials, it will block error signals that would otherwise yield positive associations between E and the dig response (Fig. 7).

In addition to this basic blocking logic, there are two important implications of the fact that E is also a member of the D+E− training pair. First, this training pattern will cause the close E stimulus to be associated with the switch response, in conflict with any E-dig association derived from the E+F− case. This produces a net switch association for E in the intact network when the E-dig association is blocked. Second, this D+E− case makes the E stimulus by itself ambiguous, which under an error-driven learning rule will cause the network to rely more strongly on the hippocampal conjunctive representation of Ef, enhancing the blocking effect.

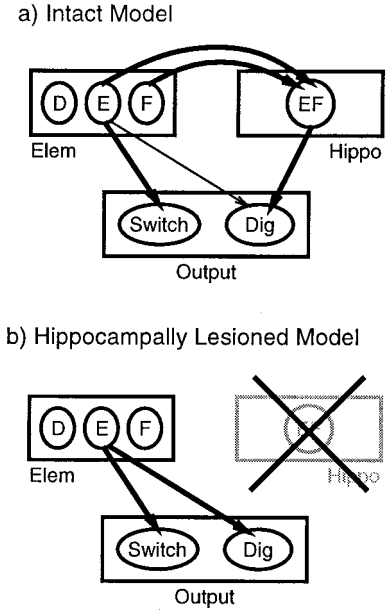

FIGURE 7. a: Illustration of how rapid hippocampal encoding of the EF training case can block the elemental associations between E and the dig output response (shown by a thinner line), causing the overall balance to favor the E-switch association, which does not suffer from this blocking effect and is therefore stronger (as shown by the thicker line). B: In the hippocampally lesioned model, this blocking effect from the hippocampus is not present, and the elemental associations between E and the dig output are roughly equal to those to the switch output. Only the critical weights between the close elemental representations are shown.

In short, learning of the association between the close E elemental representation and the dig response is blocked by the hippocampal conjunctive representation of Ef and the dig response. The lesioned model does not automatically construct the conjunctive representation of Ef, so this source of blocking is not present. The resulting differential associative weights then lead to the observed testing performance, with the intact model reliably choosing B over E and the lesioned model not.

The same reasoning can be applied to understand why the intact model has a stronger B-dig association than the lesioned model. Specifically, the hippocampus rapidly learns the A+B− problem and therefore blocks the association of b (far) with switch. This blocking then allows the B-dig association from the B+C− problem to dominate, producing the observed stronger B-dig association.

Lack of Hippocampal Blocking on Internal Problems

One remaining question needs to be addressed: if the hippocampus contributes to learning the anchor problems, why doesn't it also contribute to learning the internal problems (B+C−, C+−D−, D+E−)? This question is important because if the hippocampus participated in providing rapid conjunctive representations for the internal problems, then these representations should block the learning of associations for all of the stimuli, not just B and E. It is the differential effects of the hippocampus on these stimuli that lead to the transitive behavior of the network.

Our answer is that the internal problems are nonlinear discrimination problems. The composition of each problem overlaps completely with another problem. For example, B+C− and C+D− share C; and C+D− and D+E− share D. Therefore, each stimulus is completely ambiguous, experienced equally often with a rewarded and nonrewarded outcome. Such problems can only be solved if the animal or model constructs conjunctive representations of the stimulus pairs. So, one might expect a strong contribution from the hippocampus (cf. Sutherland and Rudy, 1989). Nevertheless, when we previously explored in detail how our model solves nonlinear discrimination problems, we found that the hippocampus does not typically contribute to faster learning (O'Reilly and Rudy, 2001). This is consistent with the literature (see Rudy and Sutherland, 1995, for review). We refer the reader to the report by O'Reilly and Rudy (2001) for a detailed discussion of why this is the case, with only a quick summary in the present study. The critical point is that in addition to encoding new memories, the hippocampus must also recall previously stored memories, which occurs via pattern completion supported by the CA3 recurrent collaterals, among other things. Pattern completion and pattern separation (which leads to new encoding of separated conjunctive representations) are fundamentally in conflict with each other (O'Reilly and McClelland, 1994); which one dominates depends on the level of input pattern overlap. Therefore, the highly overlapping internal problems will tend to trigger recall instead of new encoding. Thus, the hippocampus will be retrieving the answers to other training problems instead of learning the correct answer to the current problem, and this prevents it from producing a blocking effect.

Representational Overlap Analysis

To test further our analysis of the behavior of the network, we measured the degree of overlap of representations in the association cortex layer of the intact and lesioned models, to see if the hippocampus was influencing these representations. Specifically, we expected that the rapid learning of the Ef conjunction by the hippocampus in the intact model would cause the representation for this Ef conjunction to be separated (i.e., via hippocampal pattern separation influencing the cortex) from the representation of the individual elements E and f. Indeed, this separation is what enables the model to get the Ef training problem correct (i.e., generate a dig response) even though E by itself has a net switch association in the intact model (Fig. 6). In contrast, the lesioned model has elemental E weights that are more strongly associated with dig, so it can perform correctly using just the elemental associations for E and f. To test these ideas, we computed the Euclidean distance between the association layer representations of Ef and the elements E and f, and found that this distance was 25% greater in the intact model compared with the lesioned one. Thus, hippocampal pattern separation was influencing the cortical representations to be more separated for this conjunction.

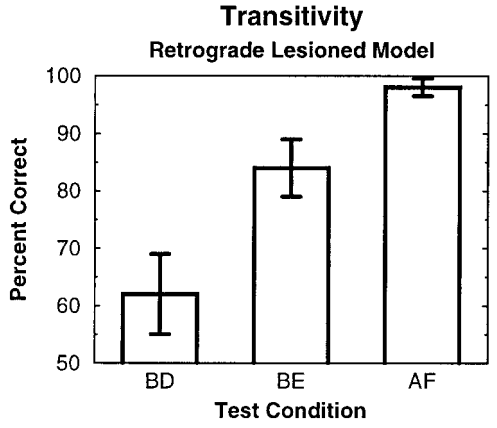

FIGURE 8. Effects of retrograde hippocampal lesions on test trial performance of the model. Note that performance is similar to that of the intact model (Fig. 5). This supports our analysis that the hippocampus is contributing during training via the blocking effect, and not via flexible retrieval during testing, as Dusek and Eichenbaum (1997) argued.

MODEL PREDICTIONS

Our simulations and analyses with the intact and lesioned models provide an explanation of the basic results reported by Van Elzakker et al. (2003) and Dusek and Eichenbaum (1997). In addition, the model generates a number of predictions that can be evaluated to further establish its validity. These predictions are based on manipulations of: when lesions are made relative to training, the training regime, and more selective lesions of hippocampal subregions.

Anterograde Versus Retrograde Effects of Hippocampal Damage

The idea that the hippocampus contributes to performance on transitivity tests by its effect on learning is fundamental to how the model explains the data. Thus, as we showed, damage to the hippocampus before training (anterograde damage) should affect test performance. Another consequence of this idea, however, is that damage to the hippocampus after training (retrograde damage) should not affect test performance. This is because the elemental associations (which determine test performance) are established during training and should be available to influence test performance even in a lesioned subject or model.

We tested this general prediction by removing the hippocampus in the model after training on the five problems but before testing it on BD, BE, and AF. Figure 8 shows that removing the hippocampal circuitry after training did not affect test performance (cf. Figs. 8 and 5).

Note that the prediction of the model is exactly the opposite of what one would derive from the representational flexibility account (Dusek and Eichenbaum, 1997). The latter view assumes that the hippocampus supports transitivity by providing a flexible retrieval strategy. Thus, it predicts retrograde damage to the hippocampus should eliminate transitivity. Therefore, a behavioral study on the retrograde effect of damage to the hippocampus on transitivity test would provide a strong test of these two views.

Replacing the Hippocampus With a Cue

According to our analysis, the hippocampus contributes to transitive inference via a blocking effect on the EF and AB problems. We reasoned that we could simulate the effects of the hippocampus in the lesioned model by simply providing an additional external cue that could produce the same kinds of blocking effects that the hippocampal conjunctive representations normally produce in the intact animal. Specifically, suppose a distinctive cue such as a shape or color (call it X) was added to the cup that contained odor F. This X cue would provide a unique, reliable basis for performance on the EF problem, in just the same way we think the conjunctive hippocampal representation does in the intact rat. In particular, this X cue would lead to rapid error reduction on this EF problem and therefore block the association of the E stimulus with the dig response, resulting in a pattern of elemental weights that should favor performance on the BE transitive inference problem. Figure 9 shows that this result was obtained in the lesioned model with the X cue provided during training. This cue was simulated by simply activating two additional units for the F stimulus (these units were never previously used in any of the other stimuli).

Improving BD Performance With a Cue

With the five-pair training problem (A+B−, B+C−, C+D−, D+E−, E+F−) studied in this and the companion article (Van Elzakker et al., 2003), performance on BD has always been worse than on BE. According to the models, this is because the associative weights to the dig response for B and D are similar. However, using the same kind of blocking logic applied in the previous prediction, we can alter the relative performance on BD and BE by adding a distinctive X cue to the E cup on D+E− trials. This manipulation causes the D association with the dig response to be blocked by the

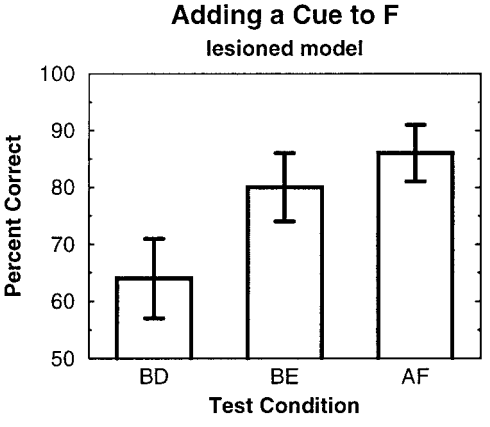

FIGURE 9. Lesioned performance of the model on test trials when a distinctive cue is added to stimulus F in training. This cue simulates the hypothesized role of the hippocampus in producing better BE test performance by blocking the association between E and dig that would otherwise be learned on the EF trials.

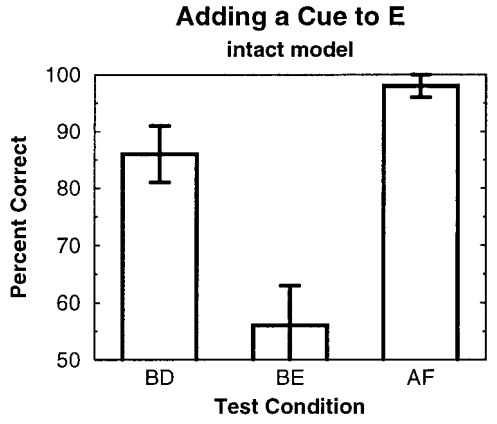

FIGURE 10. Intact performance of the model on test trials when a distinctive cue is added to stimulus E in the D+E training trials. This cue blocks the association of D with the dig response, giving D a net switch association that favors the selection of B on the BD test. Note that performance on BE is now at chance, because the distinctive cue is not added to E on E+F− trials, and the model learns that E by itself (with no cue) is rewarding.

X stimulus, giving D a net switch association. This favors the choice of B over D on the BD test, as is shown in Figure 10 (note that this effect applies to both the intact and lesioned models). Interestingly, with this distinctive X cue in place, the intact model now fails to perform well on the BE test because the X cue was not present on the E+F− trials, so E without the X cue acquired a net positive (dig) association that impaired BE performance.

CE Test Case

We can use the models to predict performance on the CE test problem, which was not tested behaviorally in rats (Van Elzakker et al., 2003). Based on the associative weights shown in Figure 6, C has a roughly neutral association with dig and switch, with just a slight dig preference. However, the intact model has a strong switch association for E, which favors selection of C on CE. Therefore, we expect that performance on CE will be better than BD, but worse than BE, because B has a stronger net dig association than C (i.e., C does not benefit from the A anchoring effect). This was confirmed (Fig. 11). CE performance in the lesioned model was worse than in the intact model and was essentially the same as BE, as one would expect.

Six-Pair Problem

Because all test probe performance in the model stems from the anchoring effects and not from "true" inference, it follows that if the network is tested on two stimuli that do not benefit directly from anchoring, it should not perform above chance. This is impossible to test in the five-pair problem explored heretofore, because all possible novel test probes involve either a B or E stimulus, which are directly affected by the anchoring phenomenon. Therefore, we extended the training stimuli to the six-pair problem going from A through G, such that the anchoring effects apply to B and F. Stimuli C, D, and E should thus be relatively unaffected, and indeed we found that testing on CE produced below-chance levels of performance in the model (Fig. 12). Conversely, the model performed well above chance for the BF test trial, as predicted by anchoring.

Effects of Overtraining

The model makes interesting predictions regarding the effects of the level of training provided to the model. In all the results presented to this point, we have provided just enough training for the network to get significantly above-chance levels of performance on the training pairs. However, an interesting dynamic occurs as we continue to train the model beyond this point: we find that the differences between the lesioned and intact models on the transitivity test items disappear. Recall that the entire difference between the intact and lesioned models lies in the blocking effect that causes the E representation to not acquire a dig association on the EF

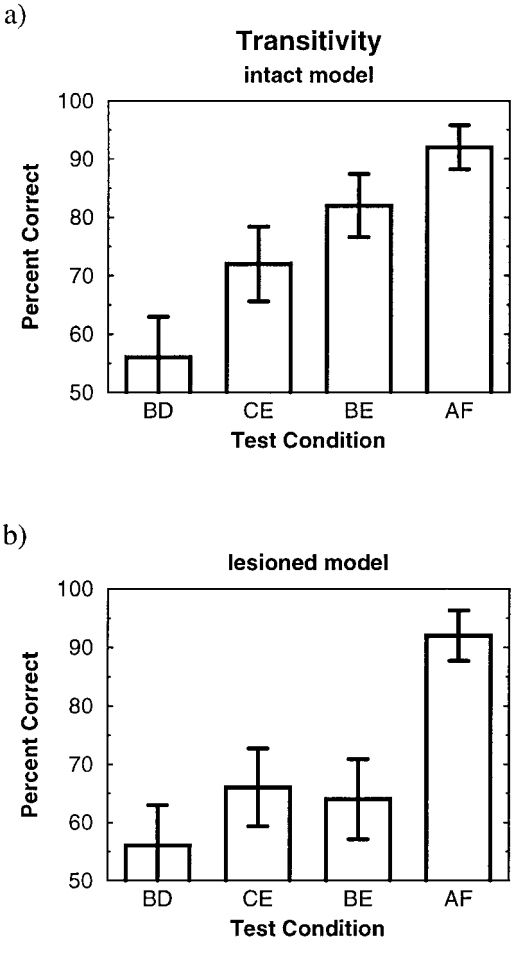

FIGURE 11. Performance on the CE test problem as compared with the previously tested problems. Based on the associative weights (Fig. 6), the E should have a net switch association in the intact model, while C is roughly neutral. Therefore, it should fall between BD (where B has a small net dig association and D is roughly neutral) and BE (where E has a net switch association, which combines with the net dig association of B to produce better performance). Lesioned model performance should be roughly the same as on BE.

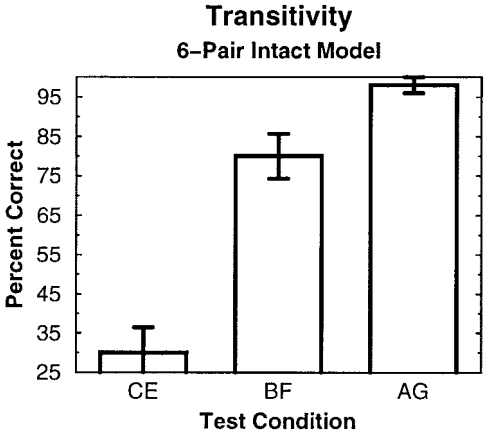

FIGURE 12. Results on test problems for the six-pair problem. Only in this problem can we test stimuli, CE, that do not benefit in any way from the anchoring effects. The poor performance on CE supports our analysis that previous good performance on other tests are due to anchoring effects.

training patterns. Instead, the E representation is associated with switch from the DE problems. If we continue to train the lesioned model, however, the EF problem is easier than the DE problem because it does have the unambiguous F stimulus. Therefore, at some point the model will learn EF quite well and stop building up a dig association for E, while still building up a switch association based on the DE problem. Therefore, we predict that BE test performance in the lesioned rat will continue to improve with training as E acquires more of a net switch association. Similar dynamics will occur for the B dig association. Figure 13 shows this prediction in the lesioned model with more extensive amounts of training.

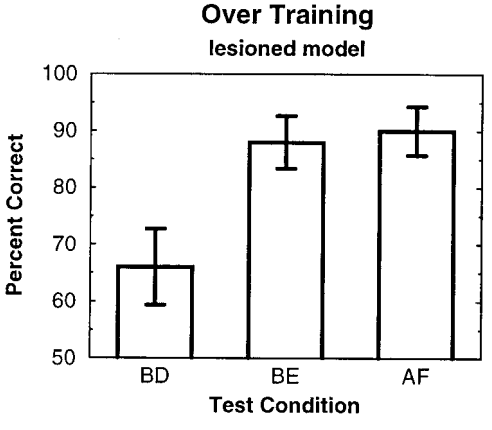

FIGURE 13. Effects of additional training on the lesioned model, which improves test performance on BE and BD. This is essentially due to a magnification of the anchoring effects, where the anchor problems become well learned and stop driving associative weights (i.e., E is less strongly associated with dig on the EF trials and B is less strongly associated with switch on the AB trials). These changes improve test performance.

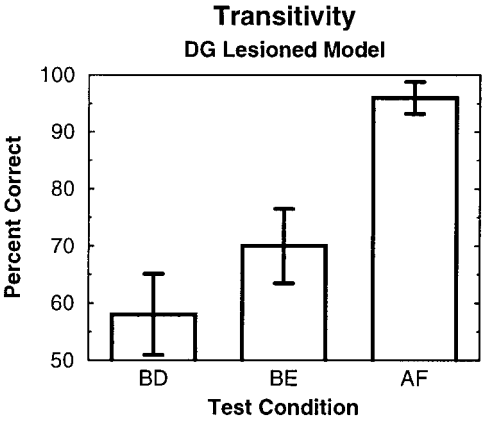

FIGURE 14. Effects of dentate gyrus (DG) lesions in the model, which produce impairments in the same direction as the entire hippocampal lesion, caused by the loss of pattern separation otherwise provided by the DG.

Lesions of Dentate Gyrus

Hippocampal pattern separation of the EF and AB training patterns is critical for the blocking effects that lead to the good test performance of the intact network relative to the lesioned one. According to our detailed analyses of the hippocampal system (O'Reilly and McClelland, 1994), the dentate gyrus (DG) is particularly important for pattern separation because of its exceptionally sparse activations and large size. Furthermore, this area is the only major area of the hippocampus that can be lesioned without completely blocking the flow of activation through the hippocampal circuit. Therefore, we tested whether DG lesions produced impairments of the same nature as those observed with lesions of the entire hippocampus. In fact, they did (Fig. 14). This provides supporting evidence for our account of the dependence on the hippocampus for transitivity. Specifically, the anatomy of the DG is instrumental in separating the necessary patterns such that the rat can choose B over E in test trials. Without the DG, pattern separation in the hippocampus is reduced, lessening the blocking effect that is otherwise seen in the intact model.

Summary of Predictions

Based on the above experiments with the models, we make the following predictions for experiments that could be tested straightforwardly in rats:

1. Retrograde hippocampal lesions (after training, before test) should produce relatively little effect compared with anterograde lesions. Note that this prediction contradicts several other findings in the literature, demonstrating that retrograde lesions have larger effects than those of anterograde lesions (e.g., Kim and Fanselow, 1992; Maren et al., 1997; Frankland et al., 1998; Richmond et al., 1999; Anagnostaras et al., 1999; Rudy et al., submitted;). This is attributed to the fact that the absence of the hippocampus can be compensated for through cortical learning (O'Reilly and Rudy, 2001), but if the hippocampus is present during training the rat will preferentially rely on it, and performance will therefore be impaired with retrograde lesions.

2. Transitive inference-like performance on the BE test case with anterograde hippocampal lesions similar to that in intact animals can be produced by training with a distinctive cue (e.g., and added color or shape) on the F stimulus. This distinctive cue simulates the effects of hippocampal conjunctive representations and produces the blocking effects that lead to choosing B over E.
3. The pattern of better BE than BD performance in intact animals can be reversed by adding a distinctive cue to the E stimulus on the DE training trials. Again, this produces blocking effects that shift the associative weights to favor B over D, but not B over E.
4. Testing on CE in intact animals should produce performance levels in between that of BD and BE. This is based on the associative weight gradients.
5. Extending to six training pairs and testing on CE in the intact animal should produce the worst test performance of any test probe. It is only in this case that neither stimulus benefits at all from anchoring effects.
6. Continued training with the lesioned animals should produce steadily better performance on BE and BD, as this effectively magnifies the anchoring effects.
7. Selective lesions of only the dentate gyrus should produce similar behavioral impairments as complete hippocampal lesions, because the pattern separation supported by this area is critical for the hippocampal contribution to normal performance.

DISCUSSION

We have used our computational model of the hippocampus and neocortex to understand how the hippocampus might contribute to tests of transitive inference performance in rats. In contrast to the ideas proposed by Dusek and Eichenbaum (1997) (see also Bunsey and Eichenbaum, 1996), who suggest that the hippocampus contributes by flexibly comparing memory traces at the point of testing, we found that the hippocampal system in our model contributed by shaping the elemental associative weights learned during training. Specifically, rapidly learned hippocampal representations produced a blocking-like effect by reducing the error signals on the anchor problems (A+B−, E+F−, in the five-pair training set), shaping the elemental associations for the B and E stimuli in a way that increases the selection of B over E on the BE transitive inference test. However, the D stimulus does not benefit from this anchoring effect, so performance on BD was near-chance. This pattern of results was observed in the behavioral study described in the companion paper (Van Elzakker et al., 2003), and is inconsistent with a logical inference account. We used our model to make predictions regarding effects of anterograde hippocampal lesions on the five-pair problem, which are consistent with the four-pair results reported by Dusek and Eichenbaum (1997). Finally, we used the model to make seven distinctive predictions that can be tested empirically, involving manipulations of training parameters (amount of training, additional cues, six-pair problem), lesions (retrograde, dentate gyrus) and other test cases (CE). Work is under way to begin testing some of these predictions in our laboratory.

We can compare the present model with an earlier incarnation (O'Reilly and Rudy, 2001). In this earlier model, which motivated the studies reported in the companion article (Van Elzakker et al., 2003), we did not incorporate an adaptive stimulus selection phase in the model, which resulted in a more equal distribution of associative weights. In this context, the only way that transitive inference could occur was through a pattern completion effect, where in the four-pair problem the BD test problem triggered pattern completion to either the BC or CD training problems. If BC was reactivated, the correct B response would be produced. However, if CD was reactivated, the associated C response was not eligible, so we imagined that this C would support the recall of BC instead, leading again to the B response. Although we do think that some of this pattern completion can be taking place in the current model, it is swamped by the larger and more reliable effects of the differential associative strength produced by anchoring effects. Thus, we see the present model as an evolution of our initial work that takes into account a critical piece that was missing from our initial analysis. The process of developing the earlier model, and deriving clear testable predictions from it, demonstrates an important strength of the computational modeling approach—it can clarify a set of ideas to the point where they can be rejected as insufficient if not supported by empirical tests. In contrast, more vague verbal theories are not always so easily testable.

Comparison With Other Theories on Transitivity

Perhaps the most direct comparison with our work can be found in the models developed by Levy and colleagues (Levy and Wu, 1997; Wu and Levy, 2001). Levy and Wu (1997) used a hippocampal-like computational model based on sequence learning that solves a similar transitive inference problem—the network chooses B over D 80% of the time. However, the only mechanism they propose is that the BD pair activates C neuronal firing which cues the selection of B, since B beats C, and D is not selected in the context of C. This interpretation is similar to the pattern completion account described above and is inconsistent with the symbolic distant effect. It is somewhat difficult, then, to understand how the same model accounts for the symbolic distance effect in subsequent research, as no mechanism is described (Wu and Levy, 2001). Furthermore, their model is purely hippocampal and so the distinctive contribution of the hippocampus relative to the cortex cannot be evaluated.

Another neural network model was developed by Siemann and Delius (1998). This model shares our focus on the associative strengths of individual stimulus items as the underlying basis for transitive behavior, but it was not related to specific neural systems. Therefore, it cannot speak to the unique contribution of the hippocampus versus cortex. Furthermore, Siemann and Delius (1998) introduced somewhat complex and specialized learning mechanisms in their model, whereas our model has shown that more generic learning mechanisms can produce differential associative strengths and transitive behavior.

We have already emphasized the contrasts between our account and that of Dusek and Eichenbaum (1997). However, there are other theoretical accounts in the literature that are much closer to our own. In particular, the anchoring theory that emerged out of our analysis of the model is similar in spirit to the value transfer theory put forth by von Fersen et al. (1991). This theory argues that associative or reward value is transferred to adjacent stimuli in the sequence. In essence, it is argued that because the rat does not have to learn much about E to approach it over F, it then only has to learn slightly more about D in order to approach it over E. In turn, this implies that the rat only has to learn slightly more about C than D, and the problem of approaching C over D is then also slightly easier. The value transfer theory implies that the effects of anchoring propagates through to all stimuli, creating a linear ordered reward relationship.

The ranked order of weight values corresponding to the elemental stimuli A through F observed in our model (Fig. 6) would seem to support the linear ordered reward relationship among the stimulus items (Trabasso and Riley, 1975; von Fersen et al., 1991). However, we postulate that the anchoring effect is enough to give a preferential benefit only to the stimuli adjacent to the anchors. Moreover, even though there is a ranked order of weights for approaching the elemental stimuli, the intact model still fails to perform well when tested on B versus D. Although the weights for approaching B are stronger than those for approaching D, the tendency to approach D still exists. If the rat happens to sample D first (i.e., D is close), it will likely approach it and fail the test. The value transfer theory seems to apply to the excitability of elemental stimuli, but the difference in value does not necessarily produce enough of an effect to cause the rat to choose the stimulus with a nominal value advantage. In short, to choose the correct stimulus consistently, the rat must have both a bias to approach it and a bias to avoid the incorrect stimulus.

The model can also address the issue of extending the training cycle by adding the inconsistent training pair of F+A−, which was implemented by Davis (1992). It was argued that this pair renders the cycle nontransitive because A is now at both the high (A>B) and low (A<F) ends of the hierarchy. We found that when this wrapping of premise pairs is implemented in the model, it takes far more trials in order to train properly. This is to be expected, because with wrapping, the anchoring conditions are effectively eliminated. There is no "crutch" to make any one pair easier to solve, as the task is completely nonlinear. Accordingly, there is no true test of transitivity, and the model performs at chance when tested on any of the test pairs that were not trained together.

SUMMARY AND CONCLUSIONS

After training on a series of discrimination problems (e.g., A+B−, B+C−, C+D−, D+E−), nonverbal organisms display transitive performance: When tested with the novel combination, BD, they choose B. It is important to understand why this phenomenon occurs because it has implications for the cognitive processes that one can assume an organism can use to adapt to its environment. One possibility is that rats, pigeons and monkeys all can extract some ordered representation from the training set (A>B>C>D>E) and display transitive performance because they use this ordered information to choose B when given the BD problem. The simpler alternative, advocated here, is that transitive performance occurs because during training each stimulus acquires some absolute amount of excitatory strength, and when the subject is faced with a novel combination (BD), it chooses the stimulus with the greatest excitatory strength.

Although organisms may have the ability to extract order from such training sequences, such a high level of cognition is not needed to produce the observed behaviors, and parsimony favors the simpler graded excitatory strength account. At the very least, to claim that behavioral transitivity is mediated by high-level relational processes, the experimenter should provide independent evidence that the choice stimuli have equal excitatory strength.

In support of the graded excitatory strength account, we demonstrated that our computational model of the hippocampus and neocortex (O'Reilly and Rudy, 2001) yields graded excitatory strengths for the individual cues and produces appropriate transitivity. Moreover, by analyzing the internal states of the network (e.g., the elemental associative weights), we were able to reconstruct how this happened and point to a subtle and perhaps transient role the hippocampus plays in contributing to the final excitatory values of the stimuli. Specifically, the ability of the hippocampus to create, both rapidly and automatically, a conjunctive representation of the EF anchor blocked excitatory conditioning to E. We were at a loss to explain this contribution before developing the computational model. Thus, we view this as example of how models can provide unique insight into subtle and complex problems in cognitive neuroscience. By testing the implications we derived from this account, we hope to determine whether this insight is correct.

REFERENCES

Anagnostaras SG, Maren S, Fanselow MS. 1999. Temporally graded retrograde amnesia of contextual fear after hippocampal damage in rats: within-subjects examination. J Neurosci 19:1106.

Bunsey M, Eichenbaum H. 1996. Conservation of hippocampal memory function in rats and humans. Nature 379:255–257.

Davis M. 1992. The role of the amygdala in conditioned fear. In: Aggleton JP, editor. The amygdala: neurobiological aspects of emotion, memory, and mental dysfunction. New York: Wiley-Liss. p 255–305.

Dusek JA, Eichenbaum H. 1997. The hippocampus and memory for orderly stimulus relations. Proc Natl Acad Sci USA 94:7109–7114.

Dusek JA, Eichenbaum H. 1998. The hippocampus and transverse patterning guided by olfactory cues. Behav Neurosci 112:762–771.

Frankland PW, Cestari V, Filipkowski RK, McDonald RJ, Silva AJ. 1998. The dorsal hippocampus is essential for context discrimination but not for contextual conditioning. Behav Neurosci 112:863–874.

Hasselmo ME. 1995. Neuromodulation and cortical function: modeling the physiological basis of behavior. Behav Brain Res 67:1–27.

Kamin LJ. 1968. "attention-like" processes in classical conditioning. In: Jones MR, editor. Miami symposium on the prediction of behavior. University of Miami Press.

Kim JJ, Fanselow MS. 1992. Modality-specific retrograde amnesia of fear. Science 256:675–677.

Levy WB, Wu X. 1997. A simple, biologically motivated neural network solves the transitive inference problem. In: Proceedings of the IEEE international conference on neural networks. Vol I. IEEE. New York, NY. p 368–371.

Maren S, Aharonov G, Fanselow MS. 1997. Neurotoxic lesions of the dorsal hippocampus and Pavlovian fear conditioning. Behav Brain Res 88:261–274.

Marr D. 1971. Simple memory: a theory for archicortex. Philos Trans R Soc Lond B 262:23–81.

McClelland JL, McNaughton BL, O'Reilly RC. 1995. Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. Psychol Rev 102:419–457.

McNaughton BL, Morris RGM. 1987. Hippocampal synaptic enhancement and information storage within a distributed memory system. Trends Neurosci 10:408–415.

O'Reilly RC. 1996. Biologically plausible error-driven learning using local activation differences: the generalized recirculation algorithm. Neural Comput 8:895–938.

O'Reilly RC. 1998. Six principles for biologically-based computational models of cortical cognition. Trends Cogn Sci 2:455–462.

O'Reilly RC, McClelland JL. 1994. Hippocampal conjunctive encoding, storage, and recall: avoiding a tradeoff. Hippocampus 4:661–682.

O'Reilly RC, Munakata Y. 2000. Computational explorations in cognitive neuroscience: understanding the mind by simulating the brain. Cambridge, MA: MIT Press.

O'Reilly RC, Rudy JW. 2001. Conjunctive representations in learning and memory: principles of cortical and hippocampal function. Psychol Rev 108:311–345.

O'Reilly RC, Norman KA, McClelland JL. 1998. A hippocampal model of recognition memory. In: Jordan MI, Kearns MJ, Solla SA, editors. Advances in neural information processing systems. Vol 10. Cambridge, MA: MIT Press. p 73–79.

Rescorla RA, Wagner AR. 1972. A theory of Pavlovian conditioning: variation in the effectiveness of reinforcement and non-reinforcement. In: Black AH, Prokasy WF, editors. Classical conditioning. II. Theory and research. New York: Appleton-Century-Crofts. p 64–99.

Richmond MA, Yee BK, Pouzet B, Veenman L, Rawlins JNP, Felden J, Bannerman DM. 1999. Dissociating context and space within the hippocampus: effects of complete, dorsal, and ventral excitotoxic hippocampal lesions on conditioned freezing and spatial learning. Behav Neurosci 113:1189–1203.

Rolls ET. 1989. Functions of neuronal networks in the hippocampus and neocortex in memory. In: Byrne JH, Berry WO, editors. Neural models of plasticity: experimental and theoretical approaches. San Diego, CA: Academic Press. p 240–265.

Rudy JW, O'Reilly RC. 2001. Conjunctive representations, the hippocampus, and contextual fear conditioning. Cognit Affect Behav Neurosci 1:66–82.

Rudy JW, Sutherland RW. 1995. Configural association theory and the hippocampal formation: an appraisal and reconfiguration. Hippocampus 5:375–389.

Rudy JW, Barrientos RM, O'Reilly RC. 2002. The hippocampal formation supports conditioning to memory of a context. Behav Neurosci (in press).

Siemann M, Delius JD. 1998. Algebraic learning and neural network models for transitive and non-transitive responding. Eur J Cogn Psychol 10:307–334.

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

Trabasso T, Riley CA. 1975. On the construction and use of representations involving linear order. In: RL. Solso, editor. Information processing and cognition: the Loyola symposium. Hillsdale, NJ: Erlbaum. p 381–410.

Van Elzakker M, O'Reilly RC, Rudy JW. 2003. Transitivity, flexibility, conjunctive representations and the hippocampus. I. An empirical analysis. Hippocampus 13:292–298.

von Fersen L, Wynne CDL, Delius JD, Staddon JER. 1991. Transitive inference in pigeons. J Exp Psychol Anim Behav Processes 17:334–341.

Wu X, Levy WB. 2001. Simulating symbolic distance effects in the transitive inference problem. Neurocomputing 30–40:1603–1610.