Psychological Review
2012, Vol. 119, No. 3, 573–616

0033-295X/12/$12.00 DOI: 10.1037/a0028681

Generalization Through the Recurrent Interaction of Episodic Memories: A Model of the Hippocampal System

Dharshan Kumaran
Stanford University and University College London

James L. McClelland
Stanford University

In this article, we present a perspective on the role of the hippocampal system in generalization, instantiated in a computational model called REMERGE (recurrency and episodic memory results in generalization). We expose a fundamental, but neglected, tension between prevailing computational theories that emphasize the function of the hippocampus in pattern separation (Marr, 1971; McClelland, McNaughton, & O'Reilly, 1995), and empirical support for its role in generalization and flexible relational memory (Cohen & Eichenbaum, 1993; Eichenbaum, 1999). Our account provides a means by which to resolve this conflict, by demonstrating that the basic representational scheme envisioned by complementary learning systems theory (McClelland et al., 1995), which relies upon orthogonalized codes in the hippocampus, is compatible with efficient generalization—as long as there is recurrence rather than unidirectional flow within the hippocampal circuit or, more widely, between the hippocampus and neocortex. We propose that recurrent similarity computation, a process that facilitates the discovery of higher-order relationships between a set of related experiences, expands the scope of classical exemplar-based models of memory (e.g., Nosofsky, 1984) and allows the hippocampus to support generalization through interactions that unfold within a dynamically created memory space.

Keywords: hippocampus, generalization, pattern separation, recurrence, complementary learning systems

Understanding the world relies upon the ability to remember both specific events from the past (e.g., where we just parked the car) and the general structure of our experiences (e.g., that dogs tend to bark; McClelland, McNaughton, & O'Reilly, 1995). These two goals, however, pose very different challenges, suggesting the need for separate neural systems based on different representational schemes (Marr, 1971; McClelland et al., 1995). Complementary learning systems (CLS) theory suggests that the hippocampus is primarily involved in the fast learning of individual episodes based on orthogonalized neural codes that minimize interference between similar experiences, leaving the task of extracting the general structure of the environment to a slow learning neocortical system that assigns overlapping codes to related concepts (McClelland et al., 1995; McClelland & Rogers, 2003; Rogers & McClelland, 2004; Rumelhart, 1990).

CLS theory (McClelland et al., 1995), building on the seminal work of David Marr and others, views the hippocampus to be a content-addressable autoassociative memory system supporting the rapid storage and retrieval of individual episodes (Burgess, 2006; Hasselmo, 1999; Hopfield, 1982; Marr, 1971; McNaughton & Morris, 1987; O'Reilly & McClelland, 1994; Treves & Rolls, 1992). A pressing computational issue considered extensively in previous work concerns how to reduce interference between similar memories stored within an episodic memory system (Burgess, 2006; Marr, 1971; McClelland et al., 1995; McNaughton & Morris, 1987; O'Reilly & McClelland, 1994; Treves & Rolls, 1992). The difficulty of this problem is captured by the following everyday example: How does one recall the specific memory of where one's car is parked today, among a myriad of highly similar competing representations? The storage in autoassociative networks of multiple similar memories with shared elements has long been known to reduce memory capacity, leading to multiple errors during recall.

One strategy that has been proposed to ameliorate this problem is *pattern separation*, whereby patterns that are similar are decorrelated (i.e., orthogonalized), thereby creating independent representations for similar experiences. By recoding patterns in this way, autoassociative networks can store and successfully recall many more similar memories without interference (Treves & Rolls, 1992). Indeed, the unique anatomical properties of the hippocampal dentate gyrus (DG), in terms of the large number of neurons present (relative to other areas such as the entorhinal cortex [ERC]), sparseness (i.e., low activity levels), and conjunctive coding, have been suggested to be ideally suited to performing pattern separation computations in the service of episodic memory

Dharshan Kumaran, Department of Psychology, Stanford University, and Institute of Cognitive Neuroscience, University College London, London, England; James L. McClelland, Department of Psychology, Stanford University.

This research was funded by a Wellcome Trust award to Dharshan Kumaran and by Air Force Research Laboratory Grant FA9550-07-1-0537 to James L. McClelland. We particularly thank Peter Dayan, Howard Eichenbaum, Ken Norman, and John Wixted for valuable discussions and comments on previous versions of the article. In addition, we thank Neal Cohen, Jeremy Glick, Robert Nosofsky, Andrew Saxe, Daniel Sternberg, Paul Thibodeau, and Anthony Wagner for helpful discussions.

Correspondence concerning this article should be addressed to Dharshan Kumaran, Institute of Cognitive Neuroscience, University College London, 17 Queen Square, London WC1N 3AR, United Kingdom; or to James L. McClelland, Department of Psychology, Stanford University, Building 420, 450 Serra Mall, Stanford, CA 94305. E-mail: d.kumaran@ucl.ac.uk or mcclelland@stanford.edu

(McClelland et al., 1995; McNaughton & Morris, 1987; Norman & O'Reilly, 2003; O'Reilly & McClelland, 1994; O'Reilly & Rudy, 2001; Treves & Rolls, 1992). As such, the DG has often been viewed to "preprocess" information for more efficient storage in CA3 during memory encoding, though how exactly the orthogonalization of patterns is maintained at the level of CA3 given the convergent nature of its inputs remains an important area of current research (Deng, Aimone, & Gage, 2010; Myers & Scharfman, 2011).

Empirical evidence suggests that the DG and CA3 subregions act to orthogonalize incoming inputs from the ERC (Kesner & Hopkins, 2006; J. K. Leutgeb, Leutgeb, Moser, & Moser, 2007; S. Leutgeb, Leutgeb, Treves, Moser, & Moser, 2004; Nakashiba et al., 2012; Vazdarjanova & Guzowski, 2004) and that this function has important behavioral consequences (Clelland et al., 2009; Deng et al., 2010; Gilbert, Kesner, & Lee, 2001; McHugh et al., 2007). Evidence from a number of studies also suggests that the putative pattern separation function of the DG is functionally relevant, with deficits in this process leading to an impaired ability to discriminate between similar stimuli or environments (Clelland et al., 2009; Deng et al., 2010; Gilbert et al., 2001; Hunsaker, Rosenberg, & Kesner, 2008; McHugh et al., 2007; Nakashiba et al., 2012). Intriguingly, the DG is also one of the few regions in the adult brain where neurogenesis (i.e., the generation of mature granule cells from neural progenitor cells) has recently been shown to occur (Deng et al., 2010). Although the exact functional and computational significance of neurogenesis is not entirely clear, recent evidence points to a role in enhancing the orthogonalization capabilities of the DG, given that its disruption produces a specific impairment in paradigms where pattern separation demands are high (Becker, 2005; Clelland et al., 2009; Deng et al., 2010; Kempermann, 2002; Nakashiba et al., 2012).

Pattern separated neural codes in the DG and CA3 regions, therefore, are viewed by a wide range of experimentalists and theoreticians as a cardinal feature of hippocampal representation and computation (J. K. Leutgeb et al., 2007; Marr, 1971; McClelland et al., 1995; McNaughton & Morris, 1987; O'Reilly & McClelland, 1994; Treves & Rolls, 1992). While optimal for episodic memory, however, distinct codes for similar input patterns come at a significant cost: In emphasizing the differences between similar memories, shared features inevitably fall by the wayside (McClelland et al., 1995). Indeed, the theoretical ideal of perfect pattern separation in the hippocampus, resulting in entirely orthogonal neural codes for highly similar experiences, would appear to offer little possibility for generalization. The idea, however, according to CLS theory, is that this apparently serious shortcoming of the scope of hippocampal processing may be relatively inconsequential, given a neocortical learning system capable of efficiently discovering the general properties, or structure, of our overall experience (McClelland et al., 1995).

Computational models of the neocortex such as the semantic network introduced by David Rumelhart (i.e., "the Rumelhart network"), operating using error correction algorithms such as backpropagation, are indeed powerful learners of the statistical structure of data sets, for instance comprised of semantic facts (e.g., a robin can fly; McClelland et al., 2010, 1995; McClelland & Rogers, 2003; Rogers & McClelland, 2004; Rumelhart, 1990). Importantly, this capacity for generalization derives from the tendency of these networks to assign overlapping internal representations for similar concepts (e.g., robin, canary), in contrast to the distinct codes associated with hippocampal processing and episodic memory (McClelland & Rogers, 2003). Neocortical networks of this form, however, are known to have several important limitations: The initial learning of new structured material must proceed slowly and be interleaved in nature (i.e., intermixing of individual training examples), leading to the very gradual incorporation of novel information. Moreover, new information that is inconsistent with prior experience must also be gradually introduced through interleaved training to avoid catastrophic forgetting of previous knowledge (French, 1999; McClelland et al., 1995; McCloskey & Cohen, 1989). In an ever-changing world, where learning opportunities are few and far between, these constraints on learning in the neocortex would indeed be severe, without cooperation from the hippocampal system and its capacity for the rapid storage and replay of individual episodes (i.e., training examples) during offline periods such as sleep (Buzsáki, 1989; Foster & Wilson, 2006; Girardeau, Benchenane, Wiener, Buzsáki, & Zugaro, 2009; Gupta, van der Meer, Touretzky, & Redish, 2010; Lee & Wilson, 2002; McClelland et al., 1995; Wilson & McNaughton, 1994). It is noteworthy, however, that new information that is consistent with prior experience may rapidly consolidate to the neocortex (e.g., in the presence of a previously learnt schema; Tse et al., 2007, 2011; also see the General Discussion).

CLS theory, therefore, leaves us able to rapidly learn the specifics of our experiences (hippocampus) and gradually appreciate the general structure of the environment (neocortex). Empirical work, however, suggests that individuals are able to exploit the relationships among items experienced in a set of related episodes after only limited exposure (e.g., within the timeframe of an experimental session; Eichenbaum, 2004; Zeithamova, Schlichting, & Preston, 2012)—a capacity that would appear to be important, both in the context of laboratory tasks such as the transitive inference paradigm and real-world settings, such as classrooms, in which one might be exposed to a set of related factual items within a single session. Indeed, the evidence further implicates the hippocampus and nearby cortical areas of the medial temporal lobe (the *hippocampal system* in our terminology) in this process, which we refer to as *rapid generalization*. This kind of rapid generalization occurs, for example, in the transitive inference task, where successful performance during probe trials (e.g., the choice of B in a B–D trial) depends on appreciating the relationship between items that have been presented in different training experiences (i.e., B+ C−, C+ D−). A role in rapid generalization, however, appears to contrast starkly with empirical and theoretical support for pattern separated neural codes in the hippocampus, which by nature tend to support episodic memory at the expense of capturing the higher order structure of a set of experiences (J. K. Leutgeb et al., 2007; Marr, 1971; McClelland et al., 1995; McNaughton & Morris, 1987; O'Reilly & McClelland, 1994; Treves & Rolls, 1992).

A fundamental tension exists, therefore, between the computational principles thought to underlie the role of the hippocampus in episodic memory (McClelland et al., 1995) and its proposed contribution to generalization (Eichenbaum, 2004; Howard, Fotedar, Datey, & Hasselmo, 2005). Here, we draw attention to this issue, which we believe has been largely neglected to date, and take steps toward developing a solution. To gain an insight into this problem, we assume a theoretical ideal of perfect pattern separation in the

hippocampus and ask how generalization can emerge under this abstract scheme. Other influential theoretical viewpoints, such as the relational theory of memory (Cohen & Eichenbaum, 1993; Eichenbaum, Dudchenko, Wood, Shapiro, & Tanila, 1999), have tended to argue that hippocampal generalization arises from the linking of related episodic experiences within a memory space. A common feature of several accounts—which we broadly refer to here as *encoding-based overlap theories*—is that the overlap between neural codes for similar episodes is critical to this process (Eichenbaum, 2004; Gluck & Myers, 1993; Howard et al., 2005; Shohamy & Wagner, 2008; also see O'Reilly & Rudy, 2001), somewhat akin to the types of neural representations instantiated in the neocortex according to CLS theory. Here, we ask whether this static overlap at the representational level is strictly necessary, given the central importance ascribed to pattern separation processes in episodic memory.

To address these issues, we introduce a model called REMERGE (recurrency and episodic memory results in generalization), which illustrates how a neural system optimized for episodic memory can also exhibit an emergent capacity for efficient generalization resulting from the dynamical interactions that occur within a recurrent system. The REMERGE model can be considered a synthesis of classical exemplar models of memory (Hintzman, 1986; Medin & Schaffer, 1978; Nosofsky, 1984; Shiffrin & Steyvers, 1997) and a class of connectionist networks, termed interactive activation and competition (IAC) models. These IAC networks, which draw on related ideas of Grossberg (1978; see also Grossberg, 1987), have been widely used in the past to characterize cognitive processes involved in perception and memory (McClelland, 1981; McClelland & Rumelhart, 1981; Nystrom & McClelland, 1992)

Overall Organization of the Article

We begin by introducing the basic architecture of the REMERGE model, relating it in broad terms to the characteristics of the hippocampal system and drawing attention to several characteristics of the model that embody key assumptions of our view concerning processing within this circuit: specifically, orthogonalized hippocampal episodic codes instantiated in our model as localist conjunctive units, recurrent processing between feature and conjunctive layers, and the co-activation during processing of multiple conjunctive units coding for related experiences. We also discuss how our approach builds on recent anatomical perspectives highlighting the potential functional significance of "big-loop" recurrency, both within the hippocampal system itself and more widely between the hippocampus and the neocortex (van Strien, Cappaert, & Witter, 2009)—an idea that contrasts with an earlier focus on the internally recurrent circuitry of the CA3 region (Burgess, 2006; Lisman, 1999; McNaughton & Morris, 1987; Treves & Rolls, 1992).

We consider three related tasks that involve rapid generalization that have been extensively used in different species, namely the transitive inference, paired associate inference, and acquired equivalence paradigms (for review, see Zeithamova et al., 2012). These experimental scenarios share a requirement for subjects to exploit information distributed over a set of related experiences and offer well-controlled environments in which to examine the mechanisms underlying generalization. We first examine the transitive inference task to bring out the core principles by which REMERGE operates. We refer to the mechanism by which REMERGE achieves generalization as recurrent similarity computation, a process by which similarity weightings are computed both for externally presented sensory inputs (i.e., as in standard exemplar models such as the generalized context model; Nosofsky, 1984) but also on feature layer inputs reconstructed by the network. We suggest REMERGE's recurrent mechanism expands the scope of classical exemplar models that rely on input-similarity-based generalization (Nosofsky, 1984) to include inferential paradigms, such as the transitivity task, while preserving the capacity of these models to perform categorization and recognition memory. We emphasize that REMERGE achieves generalization through a retrieval-based search process, rather than necessitating the formation of explicit representations of the linear hierarchy as argued by alternative accounts (e.g., Eichenbaum, 1999; Howard et al., 2005). We then provide an account of how a capacity for inference might emerge through training, based merely on the strengthening of individual premise pair memories, before turning to available empirical data, first in the transitive inference task and then the paired associate inference and acquired equivalence tasks.

We then ask whether these core principles might also provide an account of generalization-related phenomena that have recently been linked to offline periods and sleep. We relate our model to spontaneously generated neural activity in the hippocampus during resting states such as slow wave sleep (Buzsáki, 1989; Ego-Stengel & Wilson, 2010; Foster & Wilson, 2006; Girardeau et al., 2009; Gupta et al., 2010; Lee & Wilson, 2002; McClelland et al., 1995; Wilson & McNaughton, 1994). Our model predicts that replay activity in the hippocampus may be generalized in nature, reflecting a synthesis of multiple related episodic memories, rather than exclusively specific to a single episode. We present preliminary support for this hypothesis in the form of a simulation of a recent empirical study (Gupta et al., 2010), which demonstrated that a proportion of replay activity in the rodent hippocampal CA1 region consists of never-traversed shortcut sequences in a maze. We then consider the finding that a capacity for inference may emerge after an offline delay including a period of sleep in the transitivity task (Ellenbogen, Hu, Payne, Titone, & Walker, 2007).

In the final section, we relate our model of generalization to other theoretical positions on this subject. In particular, we compare our retrieval-based account of generalization to alternative treatments—the temporal context model (Howard et al., 2005), the relational theory (Cohen & Eichenbaum, 1993; Eichenbaum, 1999), and the integrative encoding hypothesis (Shohamy & Wagner, 2008), which emphasize the creation (i.e., encoding) of overlapping representations. We also examine the implications of our model for perspectives on semantic learning and, in particular, the contribution of the hippocampus and neocortex to this process. In summary, we offer recurrence in the hippocampal system as a candidate mechanism supporting generalization and flexible memory through the interactions of separate episodic codes within a dynamically changing memory space.

Basic Model Architecture

The core architecture of the recurrent model used in our simulations is illustrated in Figure 1. The model consists of two processing layers, a feature layer and a conjunctive layer, which

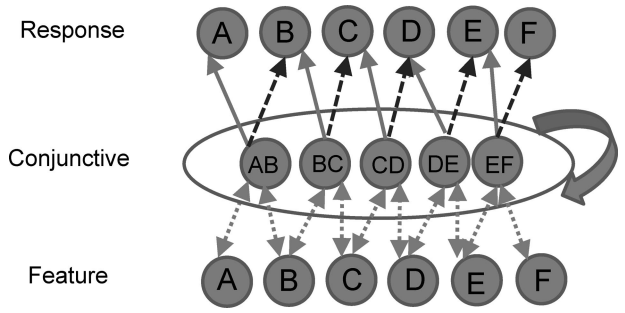

Figure 1. Schematic of model architecture, as used in transitive inference task. Basic two-layer network (feature, conjunctive) connected bidirectionally and used in all simulations. Curved arrow indicates application of hedged softmax function to the conjunctive layer, which includes the C parameter regulating overall activity. Response layer included in model architecture in simulations of transitive inference, paired associate inference, and acquired equivalence tasks. For the transitive inference task, units in feature layers denote stimuli A–F; units in conjunctive layer correspond to the five trained stimulus pairings (i.e., AB, BC, CD, DE, EF); units in the response layer correspond to the network's choice (i.e., A–F). The conjunctive layer connected with the response layer through unidirectional excitatory and inhibitory connections—for example, the AB conjunctive unit excited the A response unit and inhibited the B response unit. External input presented to the feature layer. For details of the model architecture used in specific simulations, see the main text.

communicate through excitatory bidirectional connections. In simulations of the transitive inference, paired associate inference, and acquired equivalence tasks, we also add a response layer to the configuration of the network. In the network, information is represented as patterns of activity over units in each layer. Individual units in the network take on activation values according to functions described below based on the weighted sum of the inputs received from other units in the network (i.e., their net input), subject to a competitive inhibition-like normalization process.

The model was initialized with the activity of all units set to zero, and the strength of the excitatory and inhibitory connections set to their appropriate values, as detailed in the relevant sections. External inputs were applied to the feature layer throughout a network run and processing continued through a series of timesteps or cycles, through an iterative constraint-style satisfaction process.

At each timestep, $t > 1$, the net input to a given unit (i) was determined by a weighted combination of the current net input (at time t) and that on the previous timestep ($t - 1$):

$$net_i(t) = \lambda * cnet_i(t) + (1 - \lambda) * cnet_i(t - 1),$$

where λ was set at 0.2 for all reported simulations, and $cnet_i(t)$ is given by

$$cnet_i(t) = \sum_{j=1}^{N} w_{ij}\, y_j(t) + estr * ext_i(t) + \varepsilon_\sigma,$$

where y_j is the activity of one of the N sending units j at time t, $estr$ is a parameter scaling the size of the external input (fixed at 0.5), $ext_i(t)$ is the external input presented to unit i, and ε_σ denotes normally distributed noise (included only in the replay simulation).

The activity of unit i within a given network layer was then calculated. The logistic function was used in the feature layer in cases where the input patterns presented can be thought of as being characterized by the presence or absence of certain stimuli or elements (e.g., in an A–B trial in the transitivity task):

$$y_i = \frac{1}{1 + e^{-net_i/\tau}}.$$

In this case, an activation (y value) close to one corresponds to the feature being present in the pattern of activation and an activation close to 0 corresponds to the feature being absent; the function has the property that as the net input becomes more positive, activation approaches 1, and as the net input approaches 0, the activation approaches 0. Here and below, τ is a temperature parameter, regulating how strongly activation y varies with the net input.

In the categorization and recognition tasks we consider, each stimulus is construed as having one of two or more possible values on each of several dimensions (e.g., color, size, etc.). In this case the *softmax* function was employed within each dimension (where N is the number of units in the layer):

$$y_i = \frac{e^{net_i/\tau}}{\sum_{i=1}^{N} e^{net_i/\tau}}.$$

Here, the distribution of activation over the units for the dimension corresponds to the state of information about the value of a stimulus on the dimension. For example, if on the color dimension the unit for red is highly active and the others are inactive, this corresponds to the information that the color is red. In both the logistic and softmax cases, intermediate values correspond to equivocation regarding the representation of the information in question (see below for discussion of a probabilistic interpretation of such equivocation).

A version of the softmax function termed the *hedged softmax* function was used in the conjunctive layer in all simulations:

$$y_i = \frac{e^{net_i/\tau}}{C^{1/\tau} + \sum_{i=1}^{N} e^{net_i/\tau}},$$

where τ is the temperature parameter, and C is a constant term that serves to regulate the total activation of units in the conjunctive layer. In the conventional softmax activation function, no C term is included, with overall levels of activity within a given network layer remaining constant (i.e., normalized to 1). The inclusion of the C term in our hedged softmax activation function can be seen as corresponding to the possibility that the current input is novel, rather than a match to an exemplar already stored in memory. It is included to allow us to capture the notion that the activity of conjunctive units encoding previous experiences (e.g., an AB trial) should be low when the current input is entirely novel, and the net input to all stored memories is therefore low. As such, the level of activity in the conjunctive layer increases with memory strength and the overall level of match of current inputs to past experience (see the Principal Characteristics of the Model section).

In simulations where a response layer was present in the network (e.g., transitive inference), a logistic activation function was used for units in this layer. Choice probabilities were then determined by applying the Luce choice rule (equivalent to the softmax function; Luce, 1959) to the activity values of the relevant units in the response layer (e.g., B and D units in the transitivity task):

$$P(1) = \frac{e^{y_1/\beta}}{e^{y_1/\beta} + e^{y_2/\beta}},$$

where $P(1)$ is the probability of choosing stimulus 1, and y_1 is the final activity of the unit relating to stimulus 1. Increasing values of the β lead to greater equivocation in simulated choice behavior.

The core setup of the REMERGE model has three free parameters: (1) network temperature τ (constant across all layers, and entering into the logistic and softmax activation rules); (2) the constant, C, a term entering into the denominator of the hedged softmax function (applicable only to the conjunctive layer); and (3) the β in the Luce choice rule.

Within a given simulation, a value of these parameters was selected, and then, in the transitivity (and related) simulations, the strength of weights in the network was allowed to vary to model differences in memory strength associated with extent of training (e.g., across training blocks in Ryan, Moses, & Villate, 2009) differences between subject groups (e.g., good vs. poor generalizers in Shohamy & Wagner, 2008; Zeithamova & Preson, 2010) or amount of offline delay (e.g., 20 min vs. 12 hr in Ellenbogen et al., 2007). Simulated data (unless otherwise stated) were generated through an informal search of the parameter space, with network performance indexed by the Luce choice ratios.

Principal Characteristics of the Model

Our model is grounded in several functional principles thought to characterize the hippocampal system. In several cases, we have adopted idealizations of these functional principles. Here, we discuss these characteristics, noting both the general principles and also the ways in which these have been idealized or simplified for tractability.

1. *Conjunctive coding idealized using localist codes.* In common with prior work, our model captures the notion that entire episodes or events are effectively conjunctions of recurring components (or features; Cohen & Eichenbaum, 1993; McClelland & Goddard, 1996; O'Reilly & Rudy, 2001; Rudy & Sutherland, 1989). Our model employs a localist coding scheme, using a single processing unit in the conjunctive layer to stand for the entire episode or event. This coding scheme, employed in previous applications of interactive activation and competition models to cognitive processes (McClelland, 1991), is expressly intended as an idealization. Our use of localist coding in the model instantiates the idea that the interference between individual episodic representations is minimized, as far as possible, through the use of non-overlapping orthogonalized codes (McClelland et al., 1995; McNaughton & Morris, 1987; O'Reilly & McClelland, 1994; Treves & Rolls, 1992). Neural representations in the hippocampus are thought to be sparse, conjunctive, and distributed in nature, with any single neuron participating in the representation of more than one entity or episode (Hinton, McClelland, & Rumelhart, 1986; McClelland et al., 1995; Plaut & McClelland, 2010; Quiroga, Kreiman, Koch, & Fried, 2008; Waydo, Kraskov, Quiroga, Fried, & Koch, 2006). On this view, evidence that single neurons in the medial temporal lobe (MTL) may show selective responses to only one stimulus, or concept, in a stimulus set is interpreted as consistent with the notion of sparse distributed representations, rather than localist representations (i.e., where a single neuron or distinct population of neurons represents only one entity), as advocated by Bowers (2009) or Page (2000). Our use of localist representations, therefore, does not constitute a rejection of the idea that the hippocampus uses a sparse distributed coding scheme. On the contrary, we remain convinced of the merits of the sparse, distributed view, but we employ a localist coding scheme for tractability as an idealized extreme form of a minimally overlapping and conjunctive code.

Although the coding of episodes (e.g., B–C) is localist in the conjunctive layer, as is the coding of individual components in the feature layer (e.g., A), the coding of episodes is distributed in the feature layer: Specifically, the pattern of activity during episode BC in the feature layer tends to overlap with that elicited during episode CD (see also Page, 2000). As such, the componential nature of the feature layer, which directly results in representational overlap between episodes that share elements in this layer, is critical to the ability of the recurrent activation process to produce generalization behavior over a set of related experiences.

2. *Recurrency between feature and conjunctive layers.* We emphasize the notion that pattern separated representations, of the nature presumed to exist within the hippocampus, will only support the full range of generalization we consider here when there is also recurrency within the system. The notion of big-loop recurrency within the hippocampal system is supported by anatomical and neurophysiological evidence, which is discussed below. In the model, recurrency is instantiated through bidirectional connections between the localist units in the feature and conjunctive layers. This aspect of the model draws its inspiration from the IAC model of McClelland (1981), a model with a similar architecture to that used here, but differing in its detailed quantitative formulation. We discuss the relationship to the IAC model more fully below.

3. *Learning as connection weight strengthening.* We take the approach that repeated experiences with the same experimental event (e.g., A–B premise pair in the transitive inference task) results in a strengthening of connection weights among neurons participating in the representation for that item. Instead of actually implementing an activation-based learning process in our models, the first experience with an item (e.g., A–B premise pair in the transitive inference task) creates a weak representation, modeled by a localist unit with very weak initial incoming and outgoing connection weights. Subsequent experiences are simply treated as strengthening these initially weak connections. A similar approach is typical of a wide range of more abstract memory models (e.g., McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997). This idealizes the learning process and finesses issues that still pose important challenges to our understanding of the process whereby memories for specific experiences are formed in the hippocampus and strengthened by practice.

4. *Inhibitory competition within the conjunctive layer.* Inhibitory competition is implemented within REMERGE through the use of the hedged softmax activation function over the conjunctive layer, with the degree of inhibitory competition determined by the temperature parameter, τ, and the regulatory parameter, C. Competi-

tive regulation of activity via inhibitory neurons is thought to be important in controlling an explosion of activity that would otherwise result from bidirectional positive connections within recurrent circuits. Inhibition in recurrent neural circuits, for example implemented in the IAC model and the k-winners-takes-all (k-WTA) scheme (O'Reilly & Munakata, 2000), regulates the total amount of activity and ensures that only the most active units or representations will prevail. The form of inhibitory interactions provided by the hedged softmax function allows regulation of the temperature parameter to control the degree to which activation is restricted to a single best-matching item in memory. When this parameter is set to a low value, one conjunctive unit tends to emerge as the sole winner over all others, whereas at higher temperatures more conjunctive units are allowed to participate. As previous work shows, the operating level of inhibition has a striking influence on the generality or specificity of network output produced (Grossberg, 1978). In the REMERGE model, therefore, and indeed the wider class of interactive activation networks (McClelland, 1981; McClelland & Rumelhart, 1981; also see Page, 2000), the appropriate level of inhibitory competition plays a critical role in allowing localist codes for several conjunctive experiences to interact and participate in distributed recurrent processing, thereby giving rise to an emergent capacity for generalization.

In the simulations that follow, network temperature is considered a free parameter in the model; this variable can be considered directly analogous to the sensitivity parameter central to exemplar-based models of categorization (e.g., the general context model (Nosofsky, 1984). As is typically the case in exemplar models, we leave open the question of the neural mechanisms involved in regulation of the temperature parameter. One possibility is that strategic control mechanisms, perhaps instantiated through prefrontal-hippocampal interactions, may set an appropriate level of network inhibition, according to task demands. Consistent with this notion, the prefrontal cortex (PFC) has been widely implicated in mediating selective attention and cognitive control, through an influence on local excitatory and inhibitory processes in lower level brain regions, which allows relevant target representations to dominate over competing alternatives (Desimone, 1998; Duncan, 2001; Miller & Cohen, 2001).

5. *Overall conjunctive layer activity (and correspondingly, activation of the hippocampus) increases with memory strength and match of active features to items in memory.* This property is naturally observed in the IAC model: In that model, as used in McClelland (1981; McClelland & Rumelhart, 1981), the overall activation at the conjunctive layer is dependent on the strengths of memory representations—captured by the magnitude of the connection weights between feature and conjunctive units—and on the goodness of match of the current input to experiences or items stored in memory. This feature is absent when the standard form of the softmax function is used for the conjunctive layer, since it normalizes activity across the conjunctive layer so that it always sums to 1. Introducing the constant C into the denominator of the softmax function allows the current version of the model to capture this feature of the IAC model, a feature we think any model of memory should capture. Such a constant is frequently used in instantiations of the generalized context model (GCM; e.g., Nosofsky et al., 2012). For us, this constant implements a constraint on activation stemming from the possibility that the current input is new, and should not really be treated as matching any of the items already stored in memory. If an input is presented that fails to generate a substantial net input to any of the items stored in memory, the item is treated as novel, and there is relatively little activation of representations of known items.

6. *Simulations of the network's ability to generalize are carried out in "recall" mode.* Based on the simplified learning model stated above, we explore the generalization capacity of the model, assuming that the network has already stored the relevant training examples. Our assessment of the network's performance is carried out, therefore, in "recall" mode, such that weights are fixed within a simulation, as in previous work (e.g., Hasselmo & Schnell, 1994; Wu & Levy, 2001). The scheme we use is broadly consistent with the notion of discretized hippocampal encoding and recall states suggested to arise as a function of theta cycle phase, or behavioral states characterized by varying neuromodulatory influences (e.g., waking vs. offline periods; Hasselmo, 1999, 2005). Nevertheless, we do not wish to exclude the possibility that memories are strengthened when recalled or that patterns of activation created through recurrency during recall or offline periods such as sleep may actually themselves be encoded as "stored generalizations" (see General Discussion).

7. *Simplification with respect to intrinsic variability.* Neural processes and the resulting cognitive processes are intrinsically variable. As such, including intrinsic variability is often necessary to provide a full account of all details of behavior, including capturing variance in both outcomes and response times on different trials with the same stimulus. We adhere to the view that processing is subject to such variability (Usher & McClelland, 2001), but for simplicity in the current model, we have conducted most of our simulations without including such variability in the activation process. In these cases, probabilistic human performance is approximately captured by applying the Luce choice rule to the output of the network activation process, as has generally been the practice in the interactive activation and competition model (McClelland, 1981) and in the GCM (Nosofsky, 1984). We do rely on such variability to model the probability distribution of different neural activity patterns that may arise during off-line replay of recent mental activity. Intrinsic variability also allows recurrent networks to conform to principles of Bayesian computation that may not be perfectly adhered to by the continuous approximation (McClelland, 1991; Mirman, Khaitan, Bolger, & McClelland, in press). Intrinsic variability, therefore, should be considered to be a part of the full theory of which our model is a simplified instantiation.

The Core of REMERGE: Recurrent Similarity Computation—A Synthesis of Exemplar Models and Interactive Activation and Competition Models

Our model draws upon and has important similarities with several existing models. Here, we discuss these relationships, stressing first that recurrent similarity computation in our model represents a synthesis between exemplar models of memory and interactive activation and competition networks. We then go on to consider relationships between our approach and explicitly Bayesian approaches to similarity-based generalization.

Relationship to Exemplar-Based Models of Memory

The softmax function used to determine the activity of units in the conjunctive layer grounds REMERGE in classical frameworks of similarity-based generalization and exemplar-based models of memory (Nosofsky, 1984; see below and the Appendix). Exemplar models have been influential in accounting for patterns of behavior observed in a range of domains from episodic memory tasks (e.g., recognition tasks) to categorization (Hintzman, 1986; Medin & Schaffer, 1978; Nosofsky, 1984; Shiffrin & Steyvers, 1997). These models assume that each experience tends to result in the creation of an individual memory trace, which from a geometric perspective can be viewed as represented by a point in a high dimensional space. As such, performance in a variety of settings, for example the assignment of a given stimulus to one of two categories in a categorization task, is thought to be supported by the retrieval of the most similar previously experienced exemplars, rather than any more abstract (e.g., prototype) representations.

The REMERGE model, based on localist coding of individual episodes in the conjunctive layer, can be considered a close cousin of exemplar-based models, though incorporating an additional principle of recurrence. In particular, the hedged softmax activation function used in the conjunctive layer produces activations of conjunctive units that are directly analogous to the normalized probe-exemplar similarity values calculated in an influential member of the exemplar model class, the GCM model (Nosofsky, 1984; see the Appendix for overview of principles of the GCM model and its relationship to REMERGE). As such, the activity of a unit on the network's conjunctive layer at each timestep can be viewed as reflecting the similarity of the corresponding premise pair (or stored exemplar) to the current pattern of activity on the feature layer. Processing in REMERGE, therefore, is closely related to that carried out by exemplar models: In both settings, generalization results from the parallel activation of multiple conjunctive traces determined by their similarity to the current input (the nature of which is regulated by the temperature or sensitivity parameter). In simulations presented in the Appendix, we demonstrate that REMERGE retains the capacity of exemplar-based models to perform categorization and recognition memory—the latter simulation provides a proof-of-principle demonstration that recurrency within the hippocampal circuit is broadly compatible with its well-established role in recognition memory.

We highlight one salient difference between our model and classical exemplar models: Specifically, REMERGE includes a recurrent mechanism that enables similarity-based inference and generalization to be performed not only on externally presented sensory inputs (i.e., test stimuli displayed on the screen) but also on feature layer inputs reconstructed by the network. This mechanism, which we refer to as *recurrent similarity computation*, is critical to the ability of REMERGE to perform successfully in settings like the B–D trial in the transitive inference task (see below)—by allowing the inherent similarity of two related experiences (e.g., B–C, D–E in the transitive inference task) to be captured, even when pairwise similarities in input-space are uninformative (and in this case equate to zero). While the utility of recurrency in "cleaning up" the output of the model during cued recall was noted in a previous exemplar model (MINERVA 2; e.g., the category name retrieved in response to presentation of the prototype stimulus), only limited explorations of this mechanism were performed, apparently due to limitations in available computational resources (Hintzman, 1986).

Relationship to Interactive Activation and Competition (IAC) Networks

While our use of the softmax activation function was motivated by a desire to emphasize the close relationship between our model and similarity based mechanisms of generalization, it is important to highlight parallels that exist between the REMERGE model and a class of IAC networks (McClelland, 1981; McClelland & Rumelhart, 1981; Nystrom & McClelland, 1992). As previously noted, IAC models have been used to capture a wide range of phenomena involving generalization, ranging from contextual effects in pronounceable non-words as well as words to conceptual learning tasks (McClelland, 1981; McClelland & Rumelhart, 1981; Nystrom & McClelland, 1992), and these models were our starting place in developing the framework presented here. In the first application of the IAC network architecture to memory (McClelland, 1981), the model was considered to have stored the attributes of a set of fictitious characters belonging to two rival gangs (Jets and Sharks). As in the current model, feature and conjunctive layers were comprised of localist units, connected by bidirectional excitatory connections. In this case, feature units denoted the properties of individuals (e.g., name, occupation, marital status, age, education), while units in the conjunctive layer represented different individuals (e.g., Lance). A key property of the network's mechanism of generalization is well illustrated by the following example: First, the relevant weights specifying the occupation of one particular individual (e.g., Lance: burglar) were temporarily deleted. Next, the network's ability to "fill in" Lance's appropriate occupation was tested by activating the Lance name unit on the feature layer. Critically, the network exhibited the phenomenon of generalization—in this case activating the burglar feature unit—through recurrent activation. Activating Lance's name unit activated the conjunctive-layer unit for Lance, and this in turn activated Lance's features (other than the burglar feature). This in turn led to the activation of conjunctive units coding for individuals sharing several attributes with Lance (i.e., Jim, John, George), and these conjunctive units in their turn sent activation to the burglar unit, given that all of these individuals share the burglar occupation property. Such generalization would not have occurred in standard versions of exemplar models (e.g., Nosofsky, 1984), since exemplar activations are determined solely by the direct similarity between the cues provided externally (in this case, the name, Lance) and items stored in memory. Since only one item in memory has the name Lance, no other items would be activated by the probe. It is the activation of other properties of Lance, as a result of recurrence, that provides the basis for generalization in this case, since it is only through the activation of these other properties that the activation of items sharing these properties can occur.

In summary, this example illustrates how recurrency in the IAC model (McClelland, 1981) allows an important extension of the generalization capacity of an exemplar-based mechanism. REMERGE builds on this feature of the IAC model, incorporating it into the overall framework of exemplar models of learning and memory. Overall, therefore, the REMERGE model can be considered to support a capacity for generalization through a process of

recurrent similarity computation, based on the essential operating principles of both exemplar-based models and interactive activation networks.

Connections of the REMERGE Model With Explicit Probabilistic Inference

There are also strong connections between the REMERGE model and Bayesian inference. We note, first of all, that the logistic and softmax functions used in REMERGE match very standard Bayesian computations, if connection weights and bias terms are set to be equal to the logarithms of well-known Bayesian quantities (related to the prior and the likelihood, respectively; see Hinton & Sejnowski, 1983; McClelland, 1998; Mirman et al., in press). We note further that a strong connection has recently been noted between exemplar models and approximate Bayesian inference (Shi, Griffiths, Feldman, & Sanborn, 2010); here, a key point is that the e^{net} term in our equations corresponds to the similarity term in exemplar models, and this in turn can be viewed as corresponding to the likelihood that the current input to an exemplar model would have been observed as a presentation of one of the exemplars stored in memory. To the extent that REMERGE corresponds to a classical exemplar model, then, it also corresponds to an approximate Bayesian computation.

There are two particular ways in which REMERGE differs from a standard Bayesian computation, however. The first difference lies in the fact that REMERGE treats the connection weights as quantities that change over time as knowledge of an item becomes stronger. Clearly, the weights in this case will not capture true likelihoods. They may, however, approximate varying degrees of knowledge of those true likelihoods (McClelland & Chappell, 1998). The second is in the use of recurrent activation, which, as stated above, allows REMERGE to extend its inferential capabilities beyond those of standard exemplar models. We believe there is a natural Bayesian interpretation of this extension, but we do not develop it here, since our primarily goal is to consider whether our model can account for patterns in data. Further explorations are needed to consider these issues more fully.

Anatomy of the Hippocampal System: Relationship to the Model

A detailed review of the hippocampal system is beyond the scope of this article (Amaral & Lavenex, 2006). We do, however, wish to draw the reader's attention to the anatomical and neurophysiological evidence for big-loop recurrency, which is critical to generalization in our model (van Strien et al., 2009). The hippocampal system has traditionally been viewed as unidirectional, with the output of the hippocampal system generally assumed to be the result of a single pass through the network. Accordingly, it is assumed that visual inputs from associational areas of the neocortex arrive in the superficial layers of the ERC are passed in sequential stages through the DG, CA3, and CA1 and subicular subregions to the deep layers of the ERC, before then being projected back to areas of the neocortex (Amaral & Lavenex, 2006). Of note, internal recurrency within the CA3 subregion, as distinct from the big-loop recurrency we emphasize, is a typical feature of hippocampal models (Marr, 1971; Treves & Rolls, 1994) and one viewed to be critical to the reinstatement of entire stored patterns of activity, coding for individual conjunctive memories as attractor states. However, this process is generally assumed to occur within an overall scheme of unidirectional input–output flow within the hippocampal system.

Our perspective on the hippocampal system places big-loop recurrency, which effects a recirculation of the output of the hippocampal system as a successive input, as central to the function of this circuit. We use a highly abstracted scheme of processing within the hippocampal system to bring out the essential capacities of the dynamical system we describe in a transparent fashion. Experimental data support the idea that layers deep within the hippocampus (e.g., dentate gyrus and the CA3 region) may play the role of our conjunctive layer (e.g., S. Leutgeb et al., 2004; McHugh et al., 2007). It is not, however, clear exactly which brain areas subserve the function of the featural input layer in our model. We view the feature layer as sustaining stable componential codes that are constructed for use by the hippocampal system by the slow learning neocortical system, possibly including the ERC (McClelland & Goddard, 1996). Previous work suggests that relatively dense (i.e., a relatively large fraction of cells active), structured representations of this nature, which preserve the similarity relations and exploit the redundancies present in input patterns, enhance the efficiency by which episodic experiences can be stored (McClelland & Goddard, 1996). These componential codes are likely to be represented in the ERC and possibly other medial temporal lobe regions but may also be sustained by recurrent loops involving many areas of the neocortex (including the prefrontal cortex).

Recent evidence emphasizes a potential for recurrency within the hippocampal system itself (Buzsáki, 1986; Kloosterman, van Haeften, & Lopes da Silva, 2004; van Strien et al., 2009). First, the superficial and deep layers of the ERC are anatomically interconnected, consistent with the correlated patterns of firing of neurons in the ERC across all layers during spatial tasks. Second, the hippocampus has been shown to project both to the deep and superficial layers of the ERC, rather than the deep layer alone as previously believed. Lastly, the deep layer of the ERC also projects back into the hippocampus, to the DG, with activation of this layer sufficient to activate the DG. Neurophysiological data provide evidence that the output of the hippocampal circuit re-enters the system, mediated by anatomical connections of this sort: Electrical stimulation of the subiculum activates the deep layers of the ERC and has been shown to result in re-entrance into the CA1 subregion (Kloosterman et al., 2004). Further, stimulation of the perforant path has been shown to elicit triple responses in the DG, particularly during resting states, suggesting the reverberation of neural activity within the hippocampal system (Buzsáki, 1989). Additionally, there also exist bi-directional connections between the ERC and other areas of the MTL cortex and other regions of the neocortical system (Kloosterman et al., 2004; van Strien et al., 2009). Together, these findings challenge a simple unidirectional model of processing within the hippocampal system and support the idea that a principle of "big loop" recurrence in the model may be implemented in the brain through recurrent activity involving all of these areas.

Experimental Paradigms That Involve Rapid Generalization

We consider three examples of what we consider rapid generalization: the transitive inference, acquired equivalence, and paired associate inference paradigms—all widely used across species

(Buckmaster, Eichenbaum, Amaral, Suzuki, & Rapp, 2004; Bunsey & Eichenbaum, 1996; Cohen & Eichenbaum, 1993; Dusek & Eichenbaum, 1997; Eichenbaum, 2004; Eichenbaum et al., 1999; Frank, Rudy, Levy, & O'Reilly, 2005; Greene, Spellman, Dusek, Eichenbaum, & Levy, 2001; Heckers, Zalesak, Weiss, Ditman, & Titone, 2004; McGonigle & Chalmers, 1977; Moses, Villate, & Ryan, 2006; Myers et al., 2003; Shohamy & Wagner, 2008; C. Smith & Squire, 2005; Van Elzakker, O'Reilly, & Rudy, 2003; for review, see Zeithamova et al., 2012). Here, generalization depends on the ability to appreciate the relationship between individual items (e.g., B and E in the transitive inference task) that have been presented within a set of related experiences. Further, in these settings, and indeed in other related tasks (Kumaran et al., 2007; Kumaran, Summerfield, Hassabis, & Maguire, 2009; R. A. Murphy, Mondragon, & Murphy, 2008; Preston, Shrager, Dudukovic, & Gabrieli, 2004), the capacity for generalization may arise after only a limited number of exposures to individual training examples, often within a single experimental session. Consistent with the notion that the slow learning rate operating within neocortical circuits is insufficient to support generalization performance under these conditions (McClelland et al., 1995), empirical evidence typically implicates the hippocampal system in these tasks (Eichenbaum, 2004; Zeithamova et al., 2012).

As outlined above, our perspective exposes an inherent, but neglected, tension between theories proposing a key role for the hippocampus in episodic memory based on pattern separated codes (Marr, 1971; McClelland et al., 1995; McNaughton & Morris, 1987; Treves & Rolls, 1992) and those arguing it makes an important contribution to generalization in these settings based on the overlap in neuronal codes for related episodes (i.e., encoding-based overlap theories; Eichenbaum et al., 1999; Howard et al., 2005). Here, we develop the hypothesis that pattern separated hippocampal representations can support generalization in the transitive inference, paired associate inference, and acquired equivalence tasks, through a process of recurrent similarity computation. In particular, we set out to show that the dynamic interaction of related episodic traces, made possible by recurrent flow within the network, is sufficient for generalization—thus avoiding the necessity for event codes to actually overlap at the representational level, a scheme that is thought to reduce the efficiency of an episodic memory system. Further, we also aim to provide a simple account of why the capacity for transitivity does not arise immediately but typically requires time, training on premise pairs, and/or offline delays—a phenomenon that has tended to evade formal descriptions at a mechanistic level.

Transitive Inference Paradigm

We first enquire whether a hippocampal system operating over pattern separated representations can support rapid generalization in the transitive inference task through an interaction of conjunctive units coding for related training episodes (e.g., BC, CD) mediated by recurrency. The term *transitive inference* refers to the ability to infer a relationship between stimuli (e.g., B, D) that have never been experienced together, based on previous learning of overlapping premise pairs (i.e., BC, CD). In the transitive inference task, subjects are rewarded for choosing one member of each premise pair over another (e.g., A+ B−, B+ C−, etc.). Typically, four or five pairs are used altogether, extending to D+ E−, or E+ F−, respectively. After successfully learning the premise pairs to a predefined criterion, subjects are tested on their ability to select the correct stimulus in an inference pair (i.e., B+ D−). Typically, transitivity performance improves as a function of the amount and nature (i.e., interleaved vs. blocked) of training given on premise pairs, and also as a function of the delay interposed between the end of premise pair training and inferential test (Ellenbogen et al., 2007; Moses et al., 2006; Ryan et al., 2009).

It is agreed that the hippocampus makes an important contribution to transitivity (cf. premise pair performance) across species (Eichenbaum, 2004; Zeithamova et al., 2012). While rats with hippocampal dysfunction and control subjects required a comparable number of trials to reach criterion on premise pairs, lesioned animals performed at chance levels on the critical B–D inference pair (Dusek & Eichenbaum, 1997). However, the neural mechanisms underlying the role of the hippocampus in transitivity remain unclear and subject to considerable debate. Two general classes of theories have been proposed to account for transitivity performance, those that fall under the umbrella of "associative linking" according to our terminology (Eichenbaum, 2004; Howard et al., 2005; McGonigle & Chalmers, 1977; O'Reilly & Rudy, 2001; Wu & Levy, 2001) and those that involve a form of value transfer (Frank, Rudy, & O'Reilly, 2003; von Fersen, Wynne, Delius, & Staddon, 1991). Associative linking theories, which can be applied to a range of inferential tasks, argue that generalization arises from interactions between adjacent episodes (e.g., BC, CD), typically through overlap in the relevant neuronal codes (see General Discussion for a description of the temporal context model; Howard et al., 2005) (Eichenbaum, 2004; McGonigle & Chalmers, 1977; O'Reilly & Rudy, 2001; Wu & Levy, 2001). Value transfer theories (Frank et al., 2003; von Fersen et al., 1991), in contrast, argue that transitivity performance results primarily from a difference in the reinforcement value of individual stimuli (e.g., B vs. D). The idea here is that the value of each stimulus is determined not only by its individual reinforcement history, which is equivalent for middle items (e.g., B, D in a five-item series) but additionally by the context in which it is rewarded. In this way, items (e.g., B) nearer the always rewarded "A" end anchor tend to accrue higher values than those nearer (e.g., D) to the never rewarded "E" item (in a four-pair series). As a result, a subject may correctly choose B over D in a four-pair series simply because it has a higher reward value, rather than as a consequence of any more elaborate associative linking mechanism.

Our aim here is not to summarize what is a vast literature on the subject of transitive inferences (for reviews, see Breslow, 1981; Delius & Siemann, 1998; Moses et al., 2006) or to provide an account that encompasses the full range of findings observed within the broad range of transitive inference studies. Indeed, it is widely agreed that multiple neural mechanisms contribute in parallel to performance in the transitive inference and, indeed, other mnemonic tasks, depending on the exact training conditions imposed and the species concerned (e.g., Moses et al., 2006; Zeithamova et al., 2012). Rather, we limit our focus to that of proposing a mechanistic basis for the contribution of an associative linking process to inference. We do this since associative linking mechanisms have broad relevance outside the specific domain of transitivity (unlike value transfer mechanisms), both in the setting of the paired associative inference and acquired equivalence tasks considered subsequently, as well as in other inferential paradigms

and semantic learning more generally. Further, the notion of associative linking naturally invokes a consideration of the role of the hippocampus to inferential behavior, given its capacity to rapidly set up memory representations for arbitrary conjunctions of individual items, of the sort encountered during premise trials.

Simulation of Transitive Infererence

Model architecture. As noted above (see Figure 1), in simulations of the transitive inference task, a third layer was included in the model (response layer), denoting the behavioral choice of the network (e.g., B vs. D). The following connections were present in the model: bidirectional excitatory connections between feature and conjunctive layers, unidirectional excitatory connections between conjunctive and response layers, and feedforward inhibition between conjunctive and response layers. A logistic activation function was used on feature and response layers. As outlined previously, the hedged softmax activation function was applied to the conjunctive layer to implement a form of inhibitory competition between individual units representing premise pairs, with the C parameter regulating the overall level of activity within the layer.

Prior to testing, the network was considered to have learned the premise pairs: for example, denoting the choice of A over B. As such, units A and B in the feature layer had bidirectional excitatory connections to the AB unit in the conjunctive layer; the AB unit had a positive unidirectional connection to the A unit in the response layer and had a negative feedforward connection to the B unit in the response layer. Of note, feedforward inhibition to the response layer was included in our simulation of the transitivity task since a particular stimulus (e.g., B) is consistently the incorrect (i.e., unrewarded) choice in a given context (i.e., when presented with A). The performance of the network was assessed in "recall" mode, with the strength of excitatory and inhibitory weighted connections fixed within a given run of the simulation. We did, however, change the strengths of positively weighted connections relevant to the coding of premise pairs between runs of the simulation, to assess the network's capacity for transitivity, at different stages of learning, as outlined previously.

Testing was implemented in the model by presenting external input to the visual layer, which was provided throughout the entire run, and allowing the network to settle over 300 timesteps. Testing conditions were designed to mimic the conditions of the actual experiment: Premise pair test trials involved the presentation of external input to, for instance, the B and C units. Inference pair trials consisted of external input to, for example, the B and D units.

Simulation of a Basic Capacity for Transitive Inference

We first illustrate the basic capacity of our recurrent network to exhibit the phenomenon of transitivity, with the aim of laying bare the mechanism of generalization—namely, the process of recurrent similarity computation. Notably, the characteristics of the sigmoidal neural transfer function (i.e., softmax rule) used to determine the activity of units in the conjunctive layer allow the functioning of our network to be interpreted within well-established frameworks of similarity computation and exemplar models (see the Appendix for an overview of the principles of exemplar models of memory and their relationship to REMERGE).

In the interests of clarity, we examine the timecourse of network activity during a B–D trial in more detail (see Figure 2, top panel). Activation of the B and D units on the feature layer spreads to the four relevant conjunctive units (AB, BC, CD, DE) over the first few cycles. In effect, the initial equivalence of the activity of these four units can be thought of as denoting that the current featural input (i.e., B, D) is equally similar to all of these previously learnt premise pairs. In a probabilistic sense, activity of a given conjunctive unit can be viewed as representing a posterior probability that the current pattern of feature activity (i.e., B, D unit activity) was generated through the distorted reoccurance of the relevant premise pair (e.g., BC). Over successive network cycles, the network *reconstructs* a new pattern of feature layer activity based on this initial similarity computation performed by the conjunctive layer, leading to the activity of the C unit dominating (over other units except those receiving direct external input), since it receives convergent input from both BC and CD units. Subsequently, this new pattern of feature layer activity (i.e., primarily over the B, C, and D units) causes the BC and CD units to dominate over the DE and AB units in the conjunctive layer: in effect, because the new reconstructed feature layer input is more similar to the BC and CD premise pairs (cf. the DE pair). Finally, the conjoint activation of the BC and CD conjunctive units drives the B unit to win over the D unit in the response layer.

The network's performance during a B–E inference trial (see Figure 2, middle panel) and a B–C premise trial (see Figure 2, bottom panel) is also illustrated. In a B–E trial, as in a B–D trial, the network's capacity for transitivity is mediated by a process of recurrent similarity computation, with the activity of the BC and DE units rising over the AB and EF units, despite the fact that the stimuli presented in the trial (i.e., B–E) share one feature in common with all four of these conjunctive units. Further, it can also be appreciated that the operation of recurrency results in the final activity of the CD unit, which has no features in common with the stimuli presented, being greater than that of the AB and EF units. It is worth noting, however, that in a B–C premise trial (see Figure 2, bottom panel), the presence of a direct match between the stimulus combination and previous experience results in dominant activity of the BC unit, with relatively little activation of other conjunctive units.

Generalization, therefore, in the model can be considered an emergent phenomenon, which arises a process of recurrent similarity computation, rather than through stored representations of the linear hierarchical task structure. As noted previously, traditional models of similarity based generalization (Hintzman, 1986; Medin & Schaffer, 1978; Nosofsky, 1984; Shepard, 1987), typically lacking a capacity for recurrency, are unable to perform inference in the transitivity and related tasks. This is the case because recurrency allows similarity-based computations to operate over not only externally presented sensory inputs (i.e., B and D in a B–D probe trial) but also over feature layer activity patterns reconstructed by the network, a function critical to the phenomenon of transitivity.

Here, we restrict the focus of the current treatment to the generalization capacities of the REMERGE model, given the clear links between processing in this network and a rich literature on similarity/exemplar models. Nevertheless, it is worth noting that an IAC network (using standard activation function, a resting level of zero, and parameters as detailed in

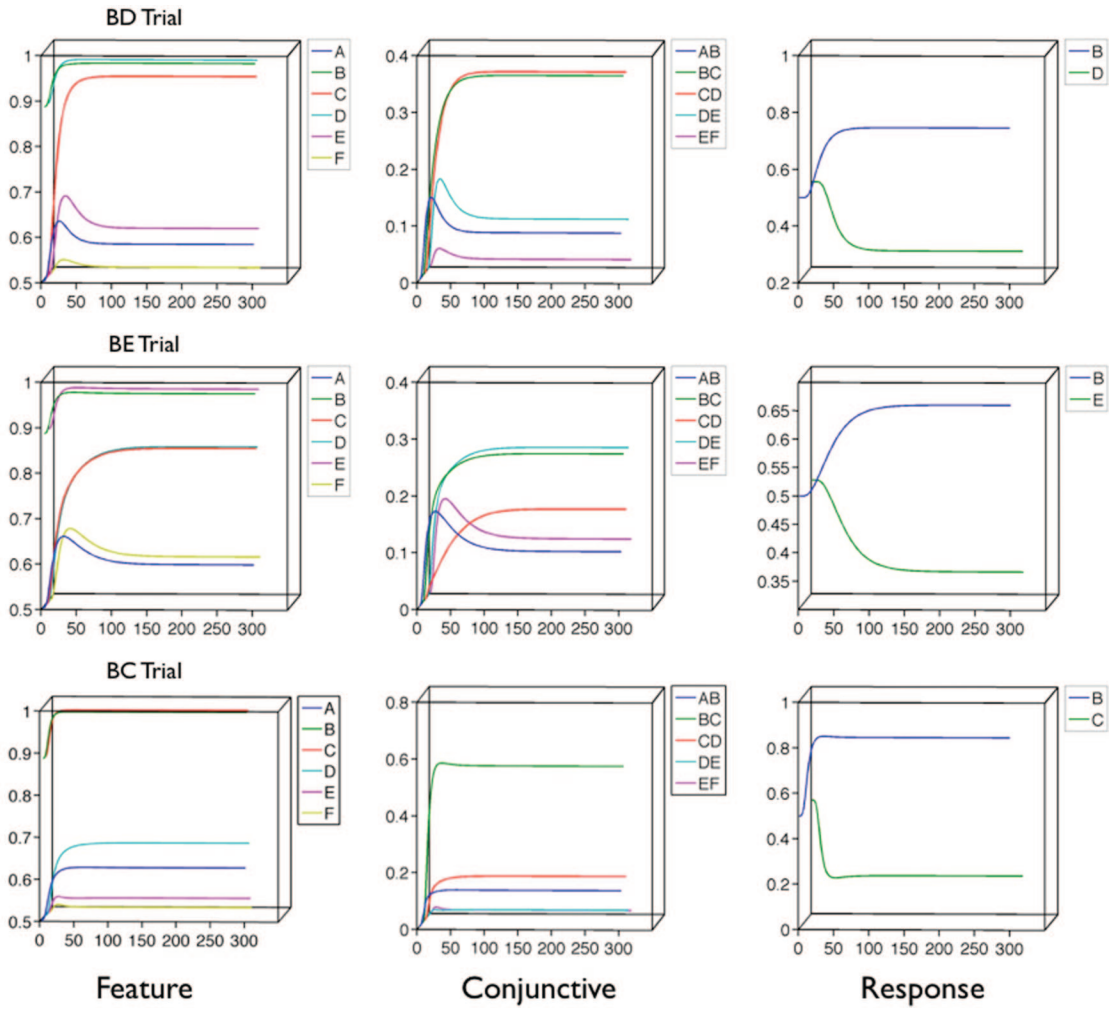

Figure 2. Illustration of the capacity of REMERGE (recurrency and episodic memory results in generalization) to exhibit the phenomenon of transitivity: simulated BD (top) and BE (middle) inference trials shown with network weights coding for premise pairs set to a value of 1.0; temperature parameter, τ, set to 0.25, C constant set to 1. Network performance during a simulated premise trial (BC) shown below with same parameters. Timecourse of network activity (*y*-axis: unit activation plotted in arbitrary units), across 300 cycles (*x*-axis). See the main text for details.

the PDP Handbook: http://www.stanford.edu/group/pdplab/pdphandbook/) is able to reproduce a basic capacity for transitivity (e.g., in a B–D trial), with the pattern of network activity observed closely mirroring that exhibited by the REMERGE model (see Figure 3). While a formal analysis of their exact relationship is likely to be complex, it would seem that generalization in an IAC network, at least in the current setting, may involve a form of recurrent similarity computation broadly analogous to that performed by the REMERGE model.

Simulating the Emergence of Inferential Capacity Through the Strengthening of Premise Pair Weights

Having demonstrated the core mechanism by which REMERGE performs transitivity, we next asked how the emergence of a capacity for inference might be simulated in the network. Our intuition was that the development of transitive responding might arise as a natural consequence of the strengthening of individual premise pair memories, implemented as an increase in the strength of the relevant weighted connections in the network. To explore this hypothesis, we adopted an idealized scheme of learning in the network (see above), whereby the first experience with an item (e.g., A–B premise pair in the transitive inference task) creates a weak representation, modeled by a localist unit with very weak initial incoming and outgoing connection weights, which is incrementally strengthened by subsequent experiences during continued training. As such, we examined the relationship between the network's performance on premise pairs and inference pairs across a range of different weight strengths.

Here, we see clearly that in the model, relatively weak connection weights can support some level of premise performance, without yet supporting generalization (see Figure 4,

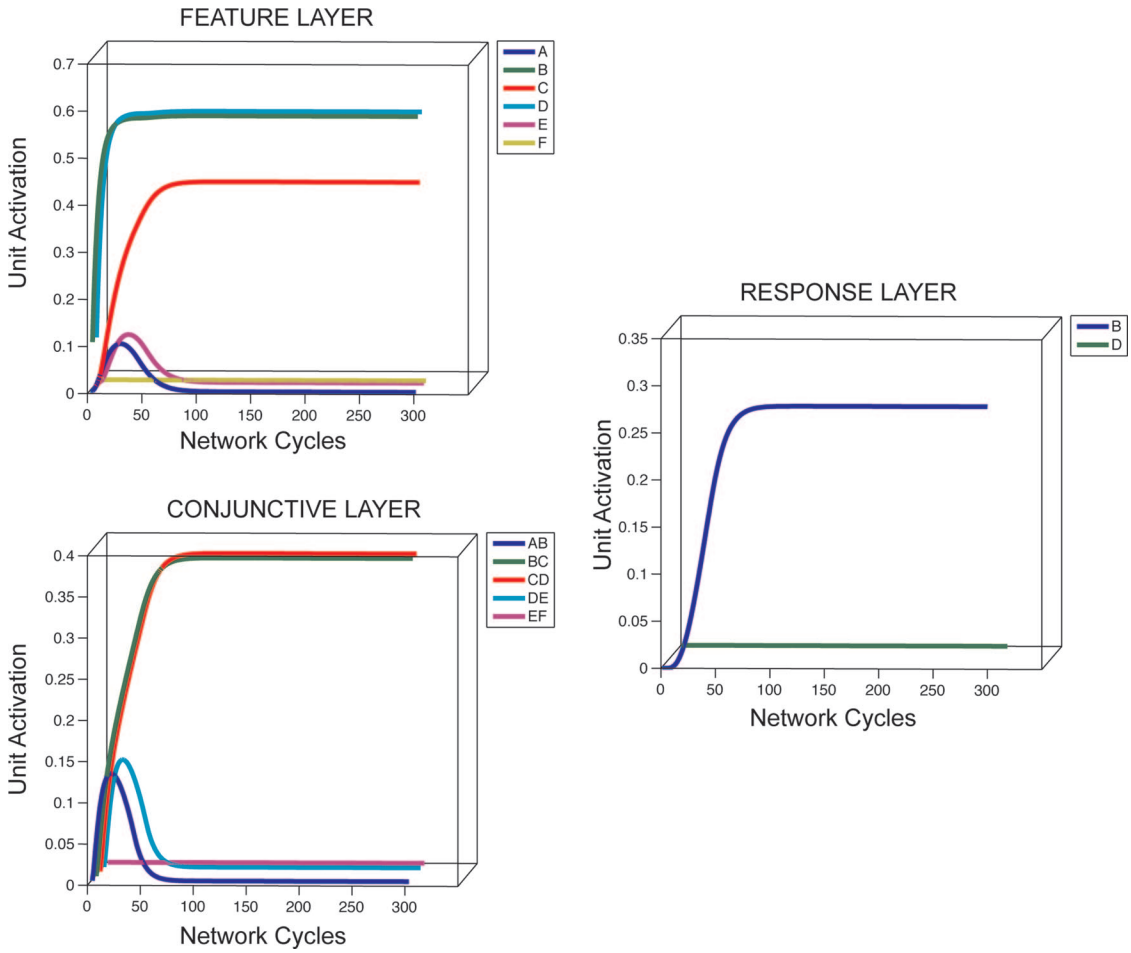

Figure 3. Illustration of the capacity of an interactive activation model (IAC) to exhibit the phenomenon of transitivity: Simulated B–D inference trial with network weights coding for premise pairs set to a value of 1.0; lateral inhibition set to –1 in conjunctive and response layers. Timecourse of network activity (*y*-axis: unit activation plotted in arbitrary units), across 300 cycles (*x*-axis). See the main text for details.

upper panel): As weights get stronger, and premise performance becomes more robust, a tendency to produce transitive inference responses emerges. Importantly, the network's failure to exhibit transitivity at lower weight strengths results from its failure (see Figure 5) to develop a graded pattern of activation over the relevant units in the feature and conjunctive layers, and in particular to exhibit greater activation of the C unit, compared to the E unit, in the feature layer during a B–D inference trial. Further, it is also evident that there is a decrement in the network's generalization performance relative to premise performance at each weight strength. The magnitude of this "lag generalization" effect is notably dependent on the setting of the temperature parameter: Lower temperature values (see Figure 4, lower panel) result in premise performance reaching near-asymptotic values before generalization performance rises significantly above chance levels, whereas higher values (e.g., Figure 4, upper panel) result in the parallel rise of premise and generalization performance, with generalization performance lagging only slightly behind premise performance.

Transitive Inference: Empirical Data and Simulation: Ryan et al. (2009)

Having demonstrated the basic capacity of REMERGE to produce transitivity, and having established the core mechanism (i.e., recurrent similarity computation) by which it achieves this, we next turned our attention to relevant empirical data. The basic phenomenon of transitivity, and its disruption by hippocampal system damage, has been observed by numerous studies (Dusek & Eichenbaum, 1997; Greene et al., 2001; Heckers et al., 2004; McGonigle & Chalmers, 1977; Moses et al., 2006; Ryan et al., 2009; C. Smith & Squire, 2005; Van Elzakker et al., 2003). Few studies (Moses et al., 2006; Ryan et al., 2009; also see Greene et al., 2001), however, have examined how the capacity for transitivity develops over the course of learning by charting the relationship between premise pair performance and inference pair performance. Given the relative paucity of suitable data, our aim in these simulations, and indeed throughout this article, was to ask whether the network can reproduce the essential features of the

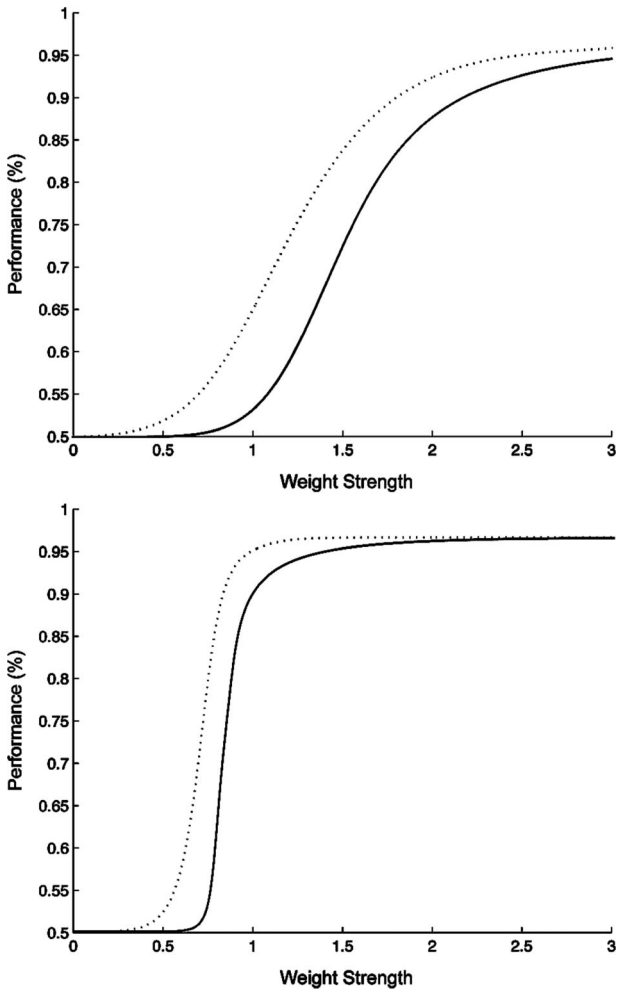

Figure 4. Simulating the emergence of a capacity for transitivity by increasing the strength of connection weights coding premise pairs. Performance (*y*-axis; Luce choice ratios expressed as a %) in premise pair trials (averaged across all pairs: dotted line), inference trials (averaged across close and distant inference pairs: solid line) plotted as a function of network weight strength. Lag generalization effect evident, whereby inferential performance at each weight strength is lower than performance on premise pairs. The magnitude of this effect is controlled by the temperature parameter, where relatively high values produce a smaller lag generalization effect (upper panel; parameters: $\tau = 0.45$; $C = 5$; $\beta = 0.3$), and relatively lower values produce a greater lag generalization effect (lower panel; parameters: $\tau = 0.2$; $C = 5$; $\beta = 0.3$).

behavioral data. Accordingly, we do not attempt to achieve a close fit to specific aspects of individual experiments through an extended search of the parameter space, but rather we consider these studies as illustrative examples representative of the extant literature. We focus on one example from the recent human literature, where the emergence of a capacity for transitivity was assessed at several timepoints over the course of training (Ryan et al., 2009).

We consider the performance of 20 young healthy volunteers who participated in the experiment by Ryan et al. (2009). The stimulus set consisted of six abstract visual pictures (A–F), yielding five premise pairs with items presented on either side of the screen, with location randomized between trials. On each premise pair trial, subjects were required to choose an item, receiving correct/incorrect feedback depending on their response. Premise pair training was conducted in five stages: In Stage 1, premise pairs were presented 10 times in consecutive order (i.e., 10 × AB, 10 × BC . . . 10 × EF). In Stage 2, premise pairs were presented five times in consecutive order (i.e., 5 × AB, 5 × BC . . . 5 × EF). Stage 3 was divided into three blocks, with premise pairs being presented three times in each block in consecutive order. During Stage 4, each premise pair was presented once in consecutive order for nine trials. In the final stage of training, premise pairs were interleaved and presented in pseudorandom order, for a total of 18 repetitions of each pair. A testing block followed each training stage in which both premise pairs (e.g., AB) and inference pairs

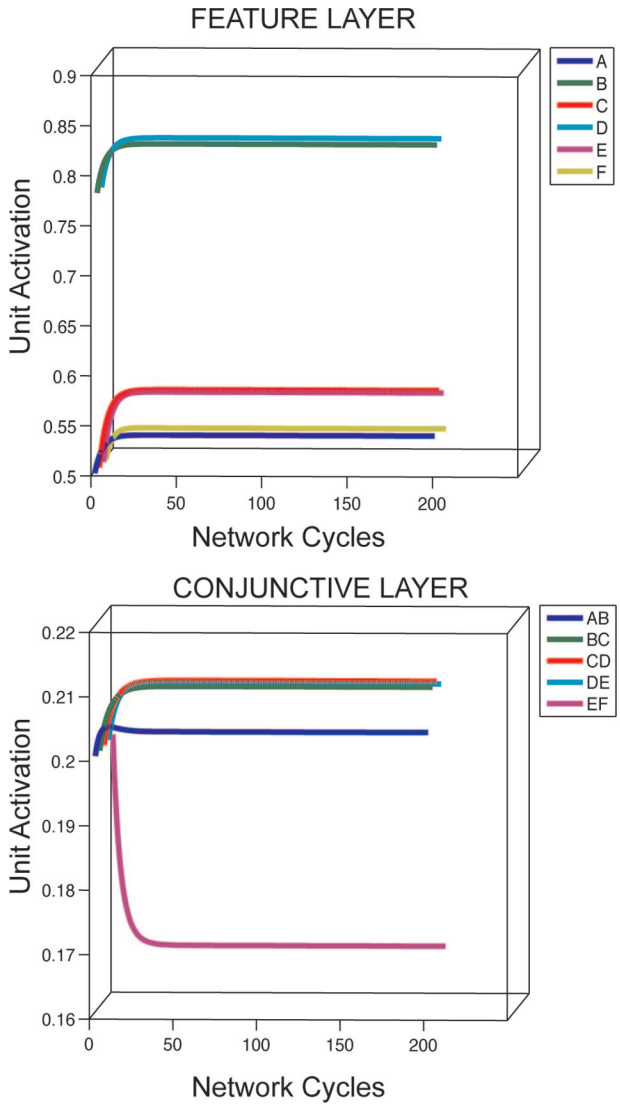

Figure 5. Transitive inference task: Illustration of failure of network to perform inference at lower weights strengths ($w = 0.3$) due to lack of graded pattern of activity over conjunctive and feature layers: $\tau = 0.25$; $C = 1$. Note similar activities of units in the conjunctive layer (cf. performance of network with $w = 1$ illustrated in Figure 2).

(e.g., BD) were presented without corrective feedback. Following the final stage of the experiment, subjects' awareness of the linear hierarchy of stimuli was assessed.

As shown in Figure 6 (upper panel), premise and inferential performance increased in parallel across the five stages of training in Ryan et al.'s (2009) study. No significant difference was observed between performance on close (i.e., B–D, C–E) and distant inference pairs in this experiment. The results of our simulation, capturing the same general pattern as seen in the data, are also illustrated in the figure (lower panel). As noted previously, the overall tendency of participants' generalization performance to lag slightly behind premise performance is reflected in the mild lag generalization effect exhibited by the model where the temperature parameter is set to a relatively high value (i.e., 0.45; see Figure 4, upper panel, for illustration of relationship between weight strength and network performance at these parameter settings).

Summary of Model Simulations

1. We demonstrate a basic capacity of the network to perform transitive inference, based on a principle of big-loop recurrence within the hippocampal system, which mediates the interaction of multiple related conjunctive units resulting in graded patterns of network activity.

2. Importantly, the network achieves inference using localist conjunctive units, designed to mirror a fundamental principle of the coding scheme employed by the hippocampus, namely pattern separation.

3. Generalization in REMERGE results from a process of recurrent similarity computation. Critically, the activity of conjunctive units can be viewed to reflect their similarity to not only externally presented sensory inputs but also inputs reconstructed by the network on the feature layer. This enables the network to exploit indirect relationships between items presented in different training experiences. In this way, recurrence can be considered to expand the explanatory power of exemplar-based models to include inferential paradigms such as the transitivity task.

4. While we focus on the REMERGE model, whose network characteristics reveal close parallels with existing exemplar based models of similarity computation, we also demonstrate that the phenomenon of transitivity is also a property of IAC networks, with processing proceeding along similar lines. We suggest, therefore, that a wider class of recurrent networks related to REMERGE is likely to display generalization capacities of the form described.

5. We show that the emergence of a capacity for transitivity can be simulated by changing a single parameter, namely the strength of weighted connections coding for premise pairs. As such, progressive training may facilitate the development of a capacity for inference (e.g., Ryan et al., 2009), merely by strengthening individual memory traces for the premise pairs themselves, rather than necessitating the discovery of new representations of the (linear) structure of the training set.

6. A characteristic feature of the model is that generalization performance tends to lag behind premise performance, as the strength of weights coding for premise pair memories is increased. Within the model, the approximately parallel rise of premise and generalization performance observed in Ryan et al.'s (2009) study is consistent with a relatively low setting of the temperature parameter, which gives rise to a similar pattern in the simulation (see Figure 4, upper panel, and Figure 6).

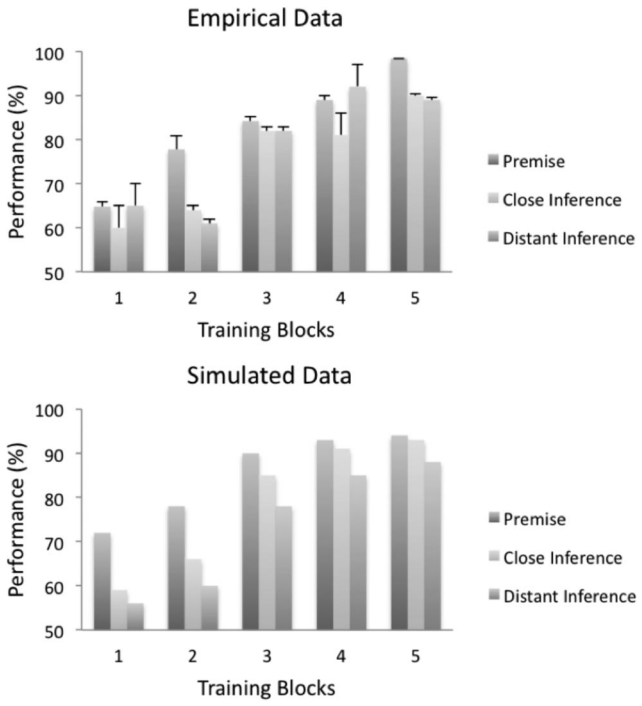

Figure 6. Empirical (upper panel) and simulated data (lower panel) relating to the young subject group of the transitive inference experiment reported in Moses et al. (2006). Performance (*y*-axis; %) in premise pair trials (averaged across all pairs: dark gray bars), close inference pairs (light gray bars: B–D and C–E trials), and distant inference trials (medium gray bars: B–E trials) across five blocks of the experiment (*x*-axis). Parameters: $\tau = 0.45$; $C = 5$; $\beta = 0.3$. Weight strengths for the five simulated blocks: 1.17, 1.32, 1.77, 2.09, 2.32. Error bars (empirical data) reflect standard errors of the mean. Model performance (expressed as % correct) derived directly from Luce choice ratios.

Paired Associate Inference and Acquired Equivalence Tasks

We next examine two tasks, the paired associative inference (PAI) and acquired equivalence (AE) paradigms, which we suggest may involve a closely related form of generalization as in the transitive inference task. Successful generalization in both the PAI and AE tasks relies on the capacity to appreciate the relationship between items presented in a set of overlapping training episodes. Performance on both these tasks, like transitivity, is disrupted by damage to the hippocampal system across species (Buckmaster et al., 2004; Bunsey & Eichenbaum, 1996; Coutureau et al., 2002; Myers et al., 2003; for review, see Zeithamova et al., 2012). One important difference between these two paradigms and the transitive inference task, however, is worth bearing in mind: The structure of the set of training examples used in the PAI and AE paradigms differs from the linear hierarchy formed by the premise pairs in the transitive inference task. Consequently, generalization in the PAI and AE paradigms cannot easily be accounted for by value transfer theory (i.e., a difference in the reinforcement value

of the individual stimuli themselves), a mechanism that may well contribute to performance in the transitive inference task (Frank et al., 2003; von Fersen et al., 1991).

In the previous section, we presented simulations illustrating how a capacity for transitive responding can emerge merely by increasing the strength of connection weights coding for premise pair memories in the network. Here, we test the validity of this account in the PAI and AE tasks. To our knowledge, empirical studies in these settings have not assessed inferential capacity at multiple timepoints over the course of premise pair training. We therefore focus on reports that subjects differ considerably in terms of their ability to generalize in both the PAI (Zeithamova & Preston, 2010) and AE (Shohamy & Wagner, 2008) tasks despite relatively uniform premise pair performance, and that "good generalizers" can be differentiated from "poor generalizers" at the neural level (functional magnetic resonance imaging [fMRI]). We ask whether the REMERGE model can provide a simple account of the difference between poor and good generalizers in terms of a difference in premise pair weight strength.

Paired Associate Inference Task

We first examined the capacity of the model to perform inference in a paradigm originally developed by Bunsey and Eichenbaum (1996) and used in a related form in a recent human imaging experiment (Zeithamova & Preston, 2010; Zeithamova et al., 2012; also see Preston et al., 2004). In the original paradigm (Bunsey & Eichenbaum, 1996), rats were first trained on premise pairs in two stages: In Stage 1, subjects were presented with cue odors (A or X) and were required to learn which of two choice odors (B or Y) was rewarded (i.e., A: B+ Y−: X: B− Y+). In the second stage, the choice odors were presented as cues, and subjects were required to choose appropriately among two new choice odors (i.e., B: C+ Z−, Y: C− Z+). Having been trained to criterion on these premise pairs, subjects then entered the generalization testing phase. In this phase, subjects were presented with cue odors from the first phase (i.e., A or X) and were required to choose between choice odors from the second training phase (i.e., C or Z) in the absence of corrective feedback. Successful generalization was indexed by the choice of C over Z in the context of cue A (i.e., A: C+ Z−) and Z over C in the context of cue X (i.e., X: C− Z+). Critically, lesions of the hippocampus produced a specific impairment in a capacity for generalization, while leaving premise pair performance relatively unaffected (i.e., no significant difference between control subjects and lesioned animals in number of trials to reach criterion on premise trials; Bunsey & Eichenbaum, 1996).

The human fMRI experiment (n = 23 subs) by Zeithamova and Preston (2010) was conducted along similar lines to these, but with some notable exceptions. Subjects intentionally learnt premise associations between overlapping pairs of objects (i.e., AB, BC, XY, YZ) in a single exposure and were not required to choose a stimulus during the study period. The entire stimulus set consisted of 144 overlapping object pairs. As in Bunsey and Eichenbaum (1996), a premise pair test trial required subjects to choose, for example, Object B over another object (e.g., Y), in the context of Object A (termed an AB premise pair trial). In generalization test trials, subjects were required to choose C over Z in the context of A (termed an AC generalization or transfer trial), and vice versa in the context of X. Successful generalization in the PAI task, therefore, involves appreciating the indirect relationship between items (e.g., A and C) presented in overlapping experiences at study (i.e., AB and BC).

Paired Associate Inference Task: Summary of Key Empirical Findings and Conclusions (Zeithamova & Preston, 2010)

1. Subjects performed well (average 82%) during the test phase of the experiment on premise pairs (i.e., AB trials; see Figure 7, top panel).

2. As a group, performance on generalization test trials (e.g., AC trials) was significantly above chance (average 64%; see Figure 7, top panel). Interestingly, however, generalization performance was found to vary greatly between individual subjects (range = 47%–90%). A binomial test was used to divide subjects into those that performed significantly above chance levels ("good" group, n = 12; mean performance = 73%) and those that did not ("poor" group, n = 11; mean performance = 54%). While performance on premise pairs was well above chance in both groups (good group: 85%, poor group: 77%), performance was significantly higher in the good group.

3. A key aim of the current study was to reveal the neural mechanisms underlying the ability of subjects to generalize in AC

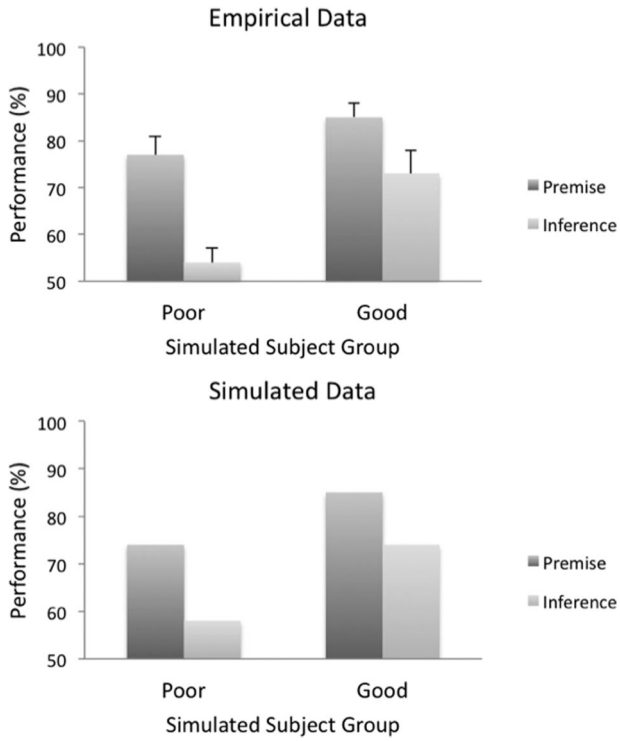

Figure 7. Empirical (upper panel) and simulated data (lower panel) relating to the behavioral findings reported in the paired associate inference experiment by Zeithamova and Preston (2010). Performance (y-axis; %) in premise pair trials (averaged across all pairs; e.g., A: B+ Y−: light gray bars), inference pairs (medium gray bars: e.g., A: C+ Z−) for "poor" and "good" subject groups (x-axis). Parameters: τ = 0.4; C = 15; β = 0.15, Weight strengths for poor and good groups, respectively: 1.47, 1.52. Error bars (empirical data) reflect standard errors of the mean. Model performance (expressed as % correct) derived directly from Luce choice ratios.

transfer trials. While the authors report findings relating to both stages of the experiment (i.e., study, test), here we focus on an observation of particular interest that relates to neural activity during the encoding (i.e., study) phase of the experiment (see Figure 8). Zeithamova and Preston (2010) performed a trial-by-trial subsequent memory analysis (Wagner et al., 1998), sorting object picture sets (e.g., AB, BC = 1 set) according to whether subjects successfully generalized during the test phase or not (i.e., AC correct vs. AC incorrect). This procedure enabled them to ask in which regions does neural activity during encoding (i.e., AB, BC trials) correlate with later successful generalization. Interestingly, they observed that activity within the hippocampus was greater during BC trials, but not during AB trials, when subjects later chose correctly (cf. incorrectly) on AC generalization trials.

The authors suggested that successful generalization was supported by integrated representations, or stored generalizations, which directly encoded the indirect relationship between items presented in related study experiences (i.e., a novel A–C association). Accordingly, the difference in generalization performance between good and poor subjects in their experiment was proposed to indicate a qualitative difference in the nature of neural representations accessible (i.e., integrated vs. premise, respectively). Based on the neural data, they further argued that the hippocampus was critical to the formation of such novel A–C associations, a process effected during BC encoding trials, and was indexed by the specific correlation observed between neural activity during BC trials and subsequent transfer success.

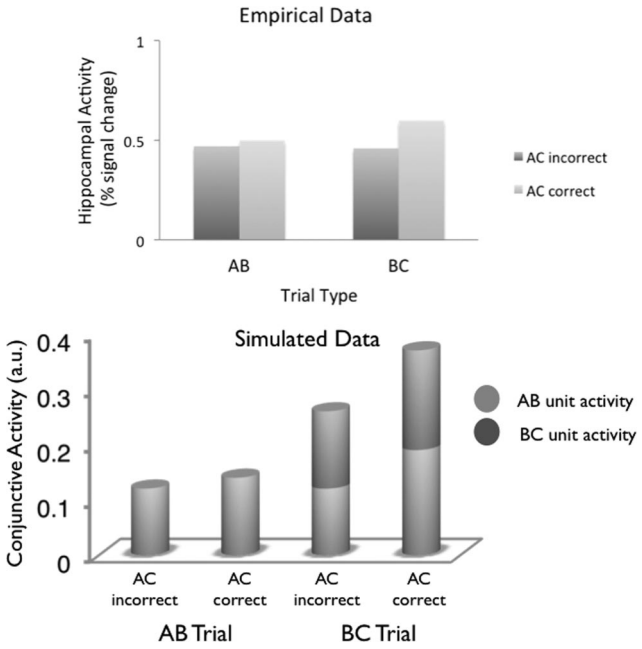

Figure 8. Empirical and simulated data relating to the functional magnetic resonance imaging (fMRI) findings observed in the paired associate inference experiment reported in Zeithamova and Preston (2010). Upper panel (empirical data): Hippocampal activity (*y*-axis: % fMRI BOLD signal change) shown for different types of encoding (premise) trials (AB and BC), as a function of whether generalization in subsequent AC transfer trial during the test phase was unsuccessful (AC incorrect: dark gray bar) or successful (AC correct: light gray bar). Significantly greater activation was observed during BC (but not AB) trials when subsequent AC trials were correct, compared to incorrect. Right panel (simulated data): Activity in the conjunctive layer (*y*-axis: arbitrary units) within the network shown for AB and BC encoding trials, at different network weight strengths (w = 1.47 and w = 1.52, for AC incorrect and AC correct trials, respectively). Light gray part of bar denotes activity of AB conjunctive unit, and dark gray part denotes activity of BC unit. Other parameters set as in behavioral simulation of poor and good performing groups—that is, $\tau = 0.4$; C = 15; $\beta = 0.15$. Data are taken from individual simulated trials, at different weight strengths. See the main text for details.

The study by Zeithamova and Preston (2010) provides a rich behavioral and neural data set indexing the emergence of successful generalization performance. As in the transitive inference task described above, we first asked whether a basic capacity for generalization can emerge from a recurrent network operating over localist codes for individual premise pairs. Next, we considered whether the behavioral performance of the good and poor generalizer groups in the current experiment could be accounted for by a strengthening of premise pair weights within the network, rather than necessarily implying a qualitative difference in the nature of neural representations in the hippocampus (i.e., integrated AC representations in the good group only). Finally, we consider the idea that the observed correlation between hippocampal activity during BC encoding trials and subsequent transfer ability may simply reflect the tendency for stronger memory traces for the individual premise pairs (e.g., AB) to favor both pattern completion (i.e., of A) during BC encoding trials and efficient generalization. As such, we aimed to link the functioning of the network to premise pair performance, generalization performance, and neural signals in the hippocampus, thereby providing a mechanistic account of the genesis of inferential behavior.

Model Specifics

As illustrated in Figure 9, the model was set up along the lines described for the transitive inference task covered previously and was designed to capture learning and generalization relating to overlapping sets of paired associates (e.g., AB, BC; XY, YZ; etc.). For the purposes of the simulation, two sets of paired associates were employed: The feature layer comprised six units denoting individual objects (A, B, C, X, Y, Z). The conjunctive layer comprised four units coding for premise pairs learnt during training (i.e., AB, BC, XY, YZ). Bidirectional excitatory connections were present between feature layer and conjunctive layer. Unidirectional excitatory connections were present between the conjunctive and response layer, the latter denoting the four objects that could be chosen during a given test trial (i.e., B, C, Y, Z). Note that feedforward inhibitory connections from the conjunctive layer to the response layer were absent in this simulation (cf. the transitive inference simulation) to mirror the specific experimental design used by Zeithamova and Preston (2010)—that is, the incorrect alternative choice in a given trial (e.g., X in a premise trial involving Objects A and B) had never previously been associated with negative feedback. In other respects (e.g., activation rules, free parameters, etc.), the model used here was directly analogous to that described above.

Prior to testing, the network was considered to have learned the premise pairs: This was coded through positive weights in the relevant connections between feature layer and conjunctive layer (e.g., A and B units in the feature layer to the AB unit in the

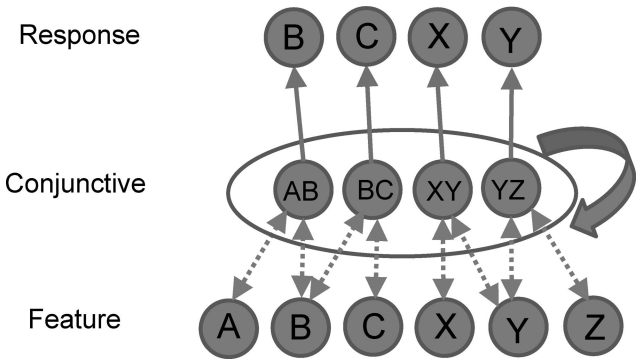

Figure 9. Schematic of model architecture, as used in the paired associative inference task. For the purposes of the simulation, two sets of paired associates were employed: The feature layer comprised six units denoting individual objects (A, B, C, X, Y, Z). The conjunctive layer comprised four units coding for premise pairs learnt during training (i.e., AB, BC, XY, YZ). Curved arrow indicates application of the hedged softmax function to this layer, including C parameter, which regulates the overall level of activity. Bidirectional excitatory connections were present between feature layer and conjunctive layer. Unidirectional excitatory connections were present between the conjunctive and response layer; the latter denoting the four objects that could be chosen during a given test trial (i.e., B, C, Y, Z). For further details, see the main text.

conjunctive layer; B and C units in the feature layer to the BC unit in the conjunctive layer) and between the conjunctive layer and response layer (e.g., AB unit in conjunctive layer to B unit in response layer; BC unit in conjunctive layer to C unit in response layer).

Testing conditions were designed to mimic the conditions of the actual experiment: As such, premise pair test trials involved presenting external input to, for instance, the A, B, and Y units on the feature layer. Generalization test trials consisted of presenting external input to, for example, the A, C, and Z feature units. Importantly, therefore, the model, and the subjects in the experiment itself, had to select between objects of equal familiarity (e.g., C and Z in a generalization trial). As previously, the behavioral performance of the network was indexed by the Luce choice ratios, determined by the final activities of units in the response layer.

As before, we examined a highly simplified version of learning in the model: The strength of the excitatory weights between feature and conjunctive layers, and conjunctive and response layers, was fixed during each network simulation but varied across simulations—that is, testing was performed with the network in recall mode. The strengthening of weights in the network coding premise pairs was designed to explore whether variation in the degree of learning could lead to differences in generalization and, therefore, simulate the observed dissociation between good and poor generalizer groups.

Simulation Results: Behavioral Data

We first examined the basic capacity of the network to behave appropriately during both premise trials (e.g., A: B+ Y−) and generalization trials (A: C+ Z−). As before, the generalization capacity of the network is mediated through graded activity patterns within the network, reflecting the interaction of conjunctive units for related premise episodes (data not shown; parameters: $\tau = 0.2$; $C = 1$; weight strength = 1). In particular, the development of patterns of activity in the network can also be viewed within a framework of recurrent similarity computation based on both externally presented input and patterns of activity reconstructed on the feature layer through pattern completion. Specifically, the externally driven activation of the A, C, and Z feature units initially results in equivalent activity of the AB, BC, and YZ conjunctive units, all of which are equally similar to the current featural input. In the next phase, this pattern of conjunctive activity reconstructs a new pattern of feature layer activity that comprises greater activity of the B unit (cf. the Y unit), driven by convergent activity from the AB and BC units. In turn, this feature layer activity pattern drives the AB and BC units to out-compete the YZ unit and, therefore, cause the C unit to win over the Z unit in the response layer.

Having established the basic capacity of the recurrent network to perform inference in this setting, we next examined its ability to account for the observed differences in behavioral performance between poor and good groups (Zeithamova & Preston, 2010), through a change in the strength of weights coding individual premise pairs (e.g., AB, BC). Empirical and simulated data are shown in Figure 7, showing that the model can indeed capture the profile of behavioral performance of poor and good groups through a difference in weight strength (i.e., 1.47 and 1.52, for poor and good groups, respectively).

Neural Data (Zeithamova & Preston, 2010)

The previous simulation suggests that the behavioral difference found at the group level in the experiment by Zeithamova and Preston (2010) may be accounted for through a quantitative change in the strength of premise pair memories. We next asked whether the subsequent memory effect reported, specifically that trial-by-trial variations in hippocampal activity during BC (but not AB) trials was correlated with subsequent performance in AC generalization test trials, could be explained in a similar fashion. Our hypothesis, therefore, was simply that better learning (i.e., greater weight strengths) of individual premise pairs (e.g., AB, BC) might underlie both successful behavioral generalization and the neural effects reported. Specifically, we reasoned that an incorrect choice on a particular AC test trial may reflect weaker memories of the relevant individual premise pairs (AB, AC). In contrast, correct AC performance may arise from stronger (i.e., weight strength = 1.52) premise memories. To test this hypothesis, we wished to examine overall levels of activity in the conjunctive layer during simulated AB and BC study trials, at weight strengths known to support either good or poor generalization ability (based on the simulation results discussed above).

Simulation Results

We therefore examined overall levels of network activity during simulated AB and BC encoding trials, at these different weight strengths (1.47 vs. 1.52). It is worth noting several points with regards to this simulation: In a simulated AB trial, we assume that the feedforward connections from the A and B feature units to the AB conjunctive unit are present (at a weight strength of either 1.47 or 1.52), but the feedback connections are absent. Given the BC

pair had not yet been experienced during an AB trial, the relevant connections were absent in the network. Likewise, at the time of a BC trial, we assume that both feedforward and feedback connections link the A and B units on the feature layer to the AB conjunctive unit; however, only feedforward connections link the B and C units on the feature layer to the BC conjunctive unit. We relate overall activity (i.e., summed) across the conjunctive layer to the hippocampal fMRI effects observed, in line with our view of the relationship between the layers of our model and the anatomy of the hippocampus.

Simulated data are illustrated in Figure 8 (lower panel). Critically, the difference in overall activity levels in the conjunctive layer during encoding trials, as a function of subsequent performance on AC transfer trials (i.e., AC correct vs. incorrect), is considerably greater during BC trials compared to AB trials. These results show that it is possible to account for the observed correlation between hippocampal activity during a BC trial and subsequent performance in AC generalization test trials simply through the modulation of the strength of individual premise pairs (e.g., AB, BC). Further inspection of the simulated data suggests that increased hippocampal activity in BC trials associated with successful AC transfer is driven by activity of the AB unit (see Figure 8, lower panel)—in line with the view of Zeithamova and Preston (2010) that the process of pattern completion (i.e., the reactivation of the previous related AB experience) makes a significant contribution to the increased hippocampal activity observed under these circumstances. It should be noted, however, that Zeithamova and Preston further argued that the simultaneous activation of the AB and BC units may naturally lead to the formation of an integrated memory trace. While we would not wish to exclude this possibility—and consider the possible contribution of stored generalizations to inference in the General Discussion—our simulations suggest that a parsimonious mechanism based on the differential strength of memories for the studied experiences may be sufficient to account for the empirical data.

Acquired Equivalence Task

We next considered the phenomenon of acquired equivalence (AE), a classic paradigm involving generalization, which bears similarities to both the transitive inference and PAI tasks. In a typical AE experiment, illustrated by a recent experiment by Shohamy and Wagner (2008), an individual learns that two stimuli (e.g., faces F1 and F2) are functionally equivalent, in that they share an association with the same outcome (e.g., scene S1). One of the stimuli (F1) is also associated during training with another scene (S2), though the other is not (F2). Critically, during inference trials in the test phase of the experiment, subjects tend to choose scene S2, over another scene (e.g., S4), when confronted with face F2, based on the functional equivalence of F1 and F2 developed during training. Importantly, therefore, the successful generalization in the AE task, as in the PAI task, necessarily involves the exploitation of indirect relationships between items presented in different training experiences (e.g., faces F1 and F2).

Empirical Data: Shohamy and Wagner (2008)

This fMRI experiment followed similar lines as described above: There were, however, 12 pairs of antecedent stimuli (i.e., F1, F2/F3, F4, etc.), rather than two, as is typical in AE paradigms. During the training phase, subjects were presented with a face (F1) and were required to choose between two different scenes (S1, S3), receiving feedback for a correct choice (S1). Importantly, the incorrect scene for one face (S3) was the correct choice for a different face on another trial (e.g., F3), preventing simple stimulus-response learning. Face-scene combinations (F1–S1) were interleaved during training, with each encountered on eight occasions, allowing subjects to reach a high level of performance (~90%). During a test phase that followed training, subjects' ability to choose appropriately on premise trials and inference trials was assessed, without the provision of feedback.

Acquired Equivalence Task: Summary of Key Empirical Findings and Conclusions (Shohamy & Wagner, 2008)

1. As a group, subjects successfully generalized during the test phase of the experiment (~75% across the whole group), thereby showing the acquired equivalence phenomenon, tending to choose S2, for instance, over scene S4 when confronted with face F2 during inference trials (see Figure 10, upper left panel).

2. As in the experiment by Zeithamova and Preston (2010), subjects varied in terms of their generalization performance (from 38% to >90%), even though premise performance was similar (~90%). Critically, a significant amount of variance in between-subjects generalization performance was accounted for by differences in the time course of activation in the hippocampus (and midbrain) during learning (see Figure 10). Specifically, in the "good" group, which was based on a median split of generalization performance (mean performance > 90%), a ramping up of activity in a hippocampal-midbrain circuit was observed from early to late training (see Figure 10, lower right panel), the degree of which correlated with subsequent generalization performance (see Figure 10, lower left panel). In contrast, no such increase in hippocampal activity was observed in the "poor" group (mean generalization performance ~60%).

The authors suggest that successful generalization in the good group, though not the poor group, may be supported by integrated representations or stored generalizations reflecting the relationship between sets of individual training experiences. Based on the neural data, they further suggest that hippocampal pattern completion during premise pair (e.g., F2–S1) trials is critical to this process, operating through the retrieval of related training episodes (e.g., F1–S1) and the subsequent creation of integrated representations (e.g., linking the three elements F1, F2, and S1: F1–F2–S1).

As in previous simulations, we asked whether the differing behavioral and neural profiles observed in the good and poor generalizer subject groups could be parsimoniously accounted for by a strengthening of premise pair weights within the network, rather than necessarily implying a qualitative difference in the nature of neural representations in the hippocampus (i.e., integrated representations in the good group only).

Specifics of Model Architecture

The architecture of the model was similar to that used to simulate the transitive inference and PAI paradigms described previously.

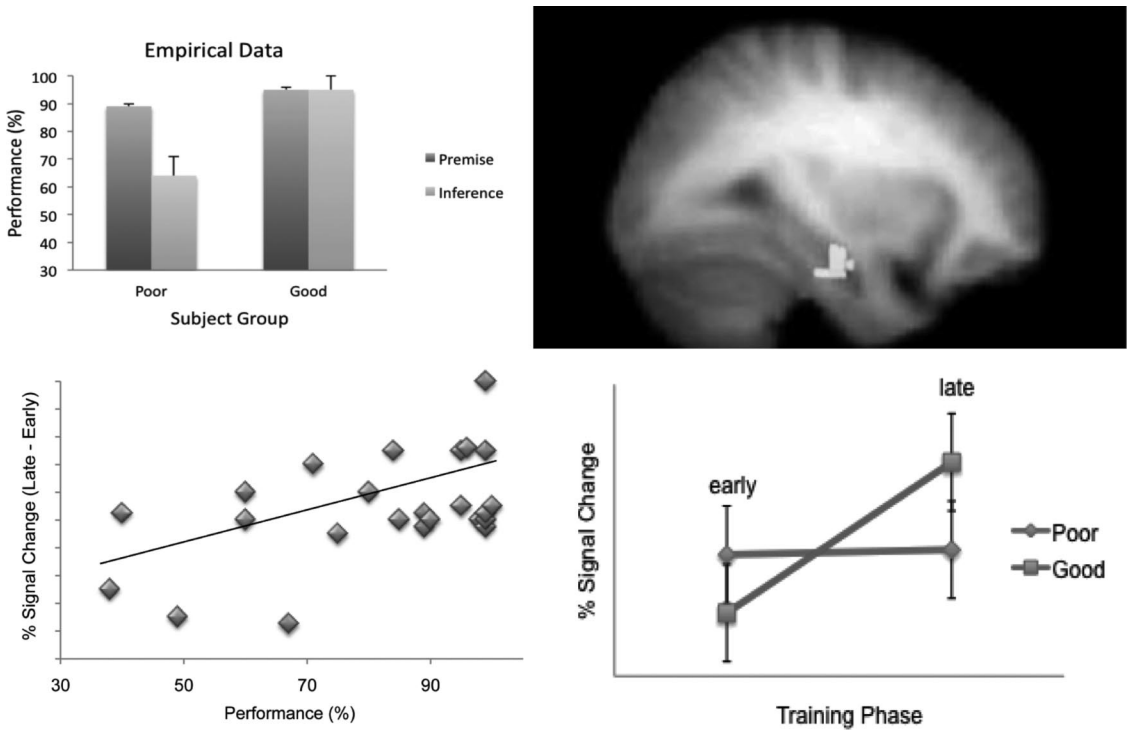

Figure 10. Acquired equivalence task: Empirical data from Shohamy and Wagner (2008). Top left panel: performance (*y*-axis; %) shown for poor (left) and good generalizer groups (right) (*x*-axis). Premise performance (dark gray bar) reflects average across all premise trials. Generalization performance (light gray bar) relates to a F2–S2 trial. While premise pair performance was near ceiling in both poor and good groups (~90%), generalization performance was far superior in the good group (~90% vs. ~60%). Bottom left panel: significant correlation (r ~0.5) between percentage change in left hippocampal BOLD signal between early and late phases of premise pair training (*y*-axis) and generalization performance (*x*-axis). Note that the actual percentage signal change magnitudes are not relevant in the current context and, therefore, are omitted for ease of interpretation. Bottom right panel illustrates that as a group, the good generalizers showed a significant increase in hippocampal activation between early and late phases of training, compared to the poor generalizers. Top right panel shows the relevant region of the left hippocampus, significant at a threshold of $p < .05$ (corrected for the volume of the hippocampus). Adapted with permission from "Integrating Memories in the Human Brain: Hippocampal-Midbrain Encoding of Overlapping Events," by D. Shohamy and A. D. Wagner, 2008, *Neuron, 60*, pp. 382 and 384. Copyright 2008 by Elsevier.

Shohamy and Wagner (2008) employed 12 pairs of face stimuli (F1, F2). We, however, used only two pairs of face stimuli (i.e., F1, F2/F3, F4) and their respective scenes (S1, S2/S3, S4) to model the relevant effects. Of note, similar results were obtained when the full stimulus set comprising 24 faces and 24 scenes was used. In the model, eight processing units in the layer correspond to stimuli F1–F4 and S1–S4; six units in conjunctive layer correspond to the six trained face-scene pairings (i.e., F1S1, F1S2, F2S1, F3S3, F3S4, F4S3); four units in the response layer correspond to the network's choice (i.e., S1–S4). The following connections were present in the model: bidirectional excitatory connections between feature and conjunctive layers; unidirectional excitatory connections between conjunctive and response layers. Of note, as in the PAI simulations, there were no inhibitory feedforward connections from conjunctive to response layer, in keeping with the structure of the paradigm used by Shohamy and Wagner (i.e., the incorrect choice in a generalization trial [e.g., S4 in an F2S2S4 trial] had never previously been associated with negative feedback).

Prior to testing, the network was considered to have learned the premise pairs: for example, denoting that scene S1 should be chosen over S3 when face F1 is present. This was coded through positive weights in the relevant connections between feature layer and conjunctive layer (e.g., F1 and S1 in the feature layer to the F1S1 unit in the conjunctive layer) and between the conjunctive layer and response layer (e.g., F1S1 unit in conjunctive layer to S1 unit in response layer). Generalization test trials, for example, consisted of presenting external input to, for example, the F2 S2 and S4 visual units. As in the PAI task, therefore, the model, and the subjects in the experiment itself, had to select between scenes of equal familiarity (e.g., S2 and S4).

As before, we explored whether the strengthening of weights in the network coding premise pairs could lead to differences in generalization and, therefore, simulate the observed behavioral and neural dissociations between good and poor generalizer groups. Simulated data at both behavioral and neural levels are shown in Figure 11. The figure shows that the network is able to account for

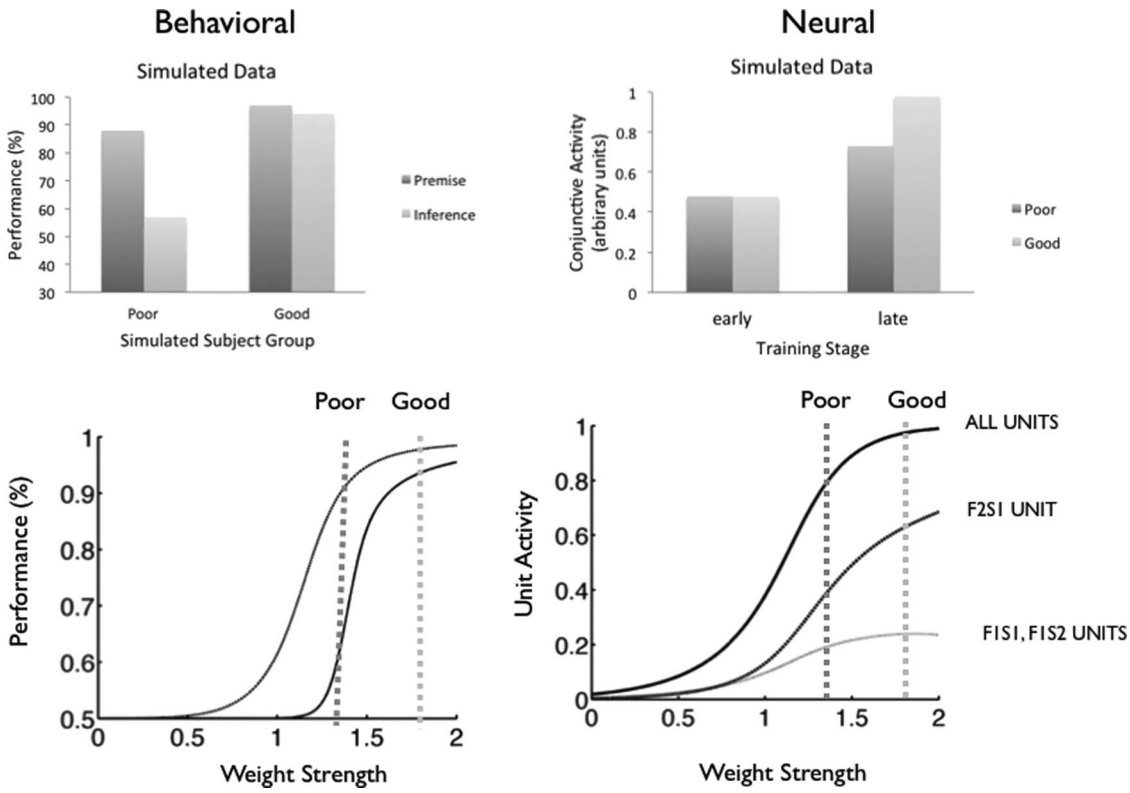

Figure 11. Acquired equivalence task: Shohamy and Wagner (2008). Simulated data. Top left panel: Simulation of behavioral data—performance (*y*-axis, Luce choice ratios expressed as a percentage) shown for poor (left) and good (right) generalizer groups (*x*-axis). Premise performance (dark gray bar) reflects average across all premise trials. While premise pair performance was significantly better in the good group, compared to the poor group, it was near ceiling in both the poor and good groups (c90%). Generalization performance (light gray bar) in contrast was far superior in the good group. Parameters: $\tau = 0.4$; $C = 10$; $\beta = 0.1$. Weight strengths for poor and good groups, respectively: 1.31, 1.82. Bottom left panel: illustration of relationship between network weight strength (*x*-axis), premise performance (light gray line; averaged across all premise pairs), and generalization performance (dark gray line). Top right panel: Neural data. Simulation of empirical finding of a greater increase in hippocampal activity between early and late phases of training (shown in bottom right panel), in good generalizer group, during F2–S1 trial. Network parameters ($\tau = 0.4$; $C = 10$) were fixed at the values used to simulate the behavioral data. Poor (dark gray bar) and good groups (light gray bar) are simulated by similar average weight strengths during the early training phase, in line with the observation that performance on premise pairs was similar in both groups (at around 70% level). In the late phase of training, good group was simulated by a network weight strength of 1.82, and poor group by a weight strength of 1.31, as in the simulation of behavioral performance. Bottom right panel: illustration of the relationship between network weight strength (*x*-axis) and activity in the conjunctive layer during an F2–S1 trial (*y*-axis; arbitrary units). Black line shows overall activity within conjunctive layer, medium gray line shows activity of the F2–S1 conjunctive unit, and light gray line shows summed activity of the F1–S1 and F1–S2 conjunctive units. Dashed vertical lines indicate simulated weight strength of the poor and good groups during late phase of training. Network parameters ($\tau = 0.4$; $C = 10$) were fixed at the values used to simulate the behavioral data. Difference in conjunctive activity between the poor and good groups during the late phase of training is due primarily to the increase in direct activation of the F2–S1 unit (see the main text for details).

the behavioral differences between poor and good generalizer groups through a difference in the strength of connection weights coding for premise pairs (see Figure 11, top right panel). Further, the network is able to reproduce the key feature of the neural differences observed, specifically the greater rise in hippocampal activity during F2–S1 trials in the good group over the course of training (see Figure 11, top right panel). However, it should be noted that the simulated poor group also shows a rise in conjunctive activity through training, a feature that differs from the profile observed in the empirical data.

It can also be seen (see Figure 11, bottom right panel) that the higher conjunctive activity levels observed in the simulated good group (cf. poor group) during F2–S1 trials in the late stage of training is primarily due to greater activity of the F2–S1 unit, which directly matches the input presented during this trial type. Indeed, in a previous simulation, we have shown that when the

stimulus combination presented during a trial directly matches a previous experience, the activity of the relevant conjunctive unit (i.e., F2–S1) dominates over all others (transitivity simulation: see Figure 2, lower panel). Our simulations, therefore, suggest that the increase in hippocampal BOLD signal observed during F2–S1 trials may not be driven primarily by the activation of related items (i.e., F1–S2, F1–S1) through pattern completion, as was the case in the simulation of the paired associate inference task (see Figure 8, lower panel). Our simulations suggest the alternative possibility that the observed increase in hippocampal activity during F2–S1 trials may be driven by activation of the directly matching conjunctive unit (i.e., F2–S1).

Summary of Paired Associate Inference and Acquired Equivalence Model Simulations and Conclusions

1. We considered the mechanism of generalization in the PAI and AE paradigms, both tasks in which the elemental values of stimuli are uninformative (cf. the transitivity task), demanding an appreciation of the higher order relationships between items presented in a set of related experiences.

2. The simulations presented provide additional support for the hypothesis that a recurrent hippocampal system, based on pattern separated neural codes and a process of recurrent similarity computation, is compatible with efficient generalization in these settings. As such, our model challenges the need for actual overlap at the representational level (Eichenbaum et al., 1999; Howard et al., 2005), a mechanism that has been argued to limit the efficiency of the hippocampus as a neural system supporting episodic memory (e.g., McClelland et al., 1995; Treves & Rolls, 1992).

3. The REMERGE model also provides a mechanistic account of why a proportion of subjects performing the PAI and AE tasks are able to generalize successfully in transfer trials, and others are not, relying simply on a difference in the strength of weights coding individual premise pairs.

4. We also suggest that our perspective provides a parsimonious mechanism by which to explain the observed relationship between trial-by-trial neural activity in the hippocampus during encoding, and subsequent generalization performance in the PAI and AE tasks. Specifically we suggest that strong premise weights lead to greater activation of either conjunctive units coding for overlapping premise experiences (i.e., AB unit during a BC trial in the PAI task) or conjunctive units that directly match the stimulus input (i.e., F2–S1 unit during a F2–S1 trial in the AE task).

5. Our model, therefore, provides a simple account of the genesis and emergence of generalization behavior, raising the question of when, and how, integrated representations (e.g., novel A–C associations in the PAI task) contribute to performance. That said, we consider in the General Discussion section how recurrency in the hippocampal system, operating during learning and over offline delays, may naturally give rise to stored generalizations of this nature.

To summarize, our perspective proposes that generalization is an emergent phenomenon that involves recurrent similarity computation mediated by the dynamic interaction of multiple stored conjunctive experiences. We also offer a simple mechanism to account for the evolution of a capacity for inference, based on the strengthening of individual premise pair memories through training. We next consider a prominent empirical finding in the literature that would appear to pose a challenge to our "weight strengthening" account of inference, specifically that patients with amnesia show relatively intact performance in classification of novel exemplars of a category, a form of generalization, even though they show markedly impaired recognition performance (Knowlton & Squire, 1993).

Preserved Generalization, but Impaired Recognition Performance in Amnesia: A Challenge for REMERGE?

In a previous section, we highlighted the important parallels between exemplar models such as GCM (Nosofsky, 1984) and REMERGE and provided evidence through simulations presented in the Appendix that REMERGE retains capacities of exemplar models in performing categorization (i.e., in the 5–4 task) and recognition memory. Here, we consider whether our model is able to capture the empirical demonstration that recognition memory performance and categorization ability may dissociate in patients with amnesia (Knowlton & Squire, 1993). We note that the interpretation of these empirical observations remains highly contentious: While a popular account views these data as supporting the operation of multiple memory systems in the brain (Knowlton & Squire, 1993), it has also been argued that a single system exemplar model, specifically the GCM (Nosofsky, 1984), may be sufficient to capture the essential features of the data (Nosofsky & Zaki, 1998). Our aim here is not to provide a resolution to this entrenched debate (Knowlton & Squire, 1993; Nosofsky & Zaki, 1998; also see Nosofsky, Little, & James, 2012) but rather to assess whether our model is able to reproduce the essential features of the reported empirical dissociation and to consider potential implications for the types of generalization supported by REMERGE.

Empirical Results: Knowlton and Squire (1993)

Twelve control subjects and 10 amnesic subjects (five with medial temporal lobe damage, five with diencephalic damage) performed a categorization task and a recognition memory task (Knowlton & Squire, 1993). During the study phase of the categorization task, subjects viewed 40 dot patterns that were high distortions of a prototype (Posner & Keele, 1968). During the test phase, they viewed four repetitions of the prototype dot pattern, 20 new low distortions, and 40 random patterns and were asked to judge whether each test pattern belonged to the same category as the patterns viewed during the study phase. In the recognition experiment, subjects viewed five random dot patterns, presented eight times each. At test, they were shown the five previously studied patterns, as well as five new random patterns, and were asked to judge whether each test pattern was old or new.

The results from the empirical study are illustrated in Figure 12 (upper panels) Two features are apparent: First, both control and amnesic subjects showed the prototypicality effect during categorization, endorsing the prototype pattern, low/high distortions, and random patterns with decreasing frequency (see Figure 12, upper right panel). Second, patients with amnesia were markedly impaired at the recognition memory test but did not perform significantly differently from control subjects at categorization (see Figure 12, upper left panel), with a significant interaction between subject group and task reported.

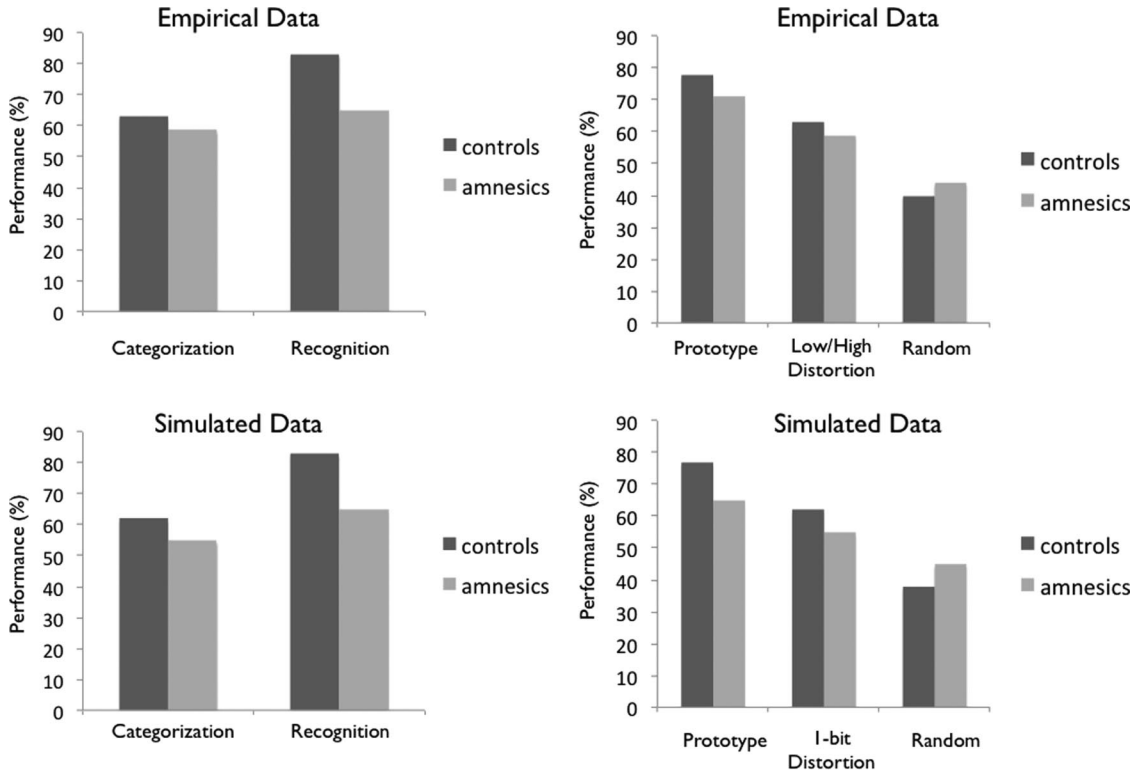

Figure 12. Empirical (upper panels) and simulated (lower panels) data relating to the experiment by Knowlton and Squire (1993). Performance (%) shown on *y*-axis and indexes probability of endorsing a test item as a category member (categorization task) or judging an item as old (recognition task). Dark gray bars = control group; light gray bars = amnesic group. Left panels show dissociation between relatively spared overall categorization performance (i.e., across all test pattern types) and impaired recognition performance. Right panels show endorsement probability for each test item type in the categorization task, illustrating prototypicality effect. Performance in the empirical study is collapsed across high and low distortions and was simulated by 1 distortion level (see the main text for details). Parameter settings for the simulation were as follows: control group (weight strength = 1.50, $C_{categorization} = 970$, $C_{recognition} = 700$) and the amnesic group (weight strength = 0.83, $C_{categorization} = 70$, $C_{recognition} = 30$), where C is the regulatory parameter entering into the hedged softmax function applied to the conjunctive layer. C was set at a level that ensured unbiased responding (i.e., equal numbers of hits and correct rejections). Temperature was fixed at 1 throughout the simulation.

Simulation

We adopt a procedure that shares principles with the application of the GCM model to this data set (see Nosofsky & Zaki, 1998, for details). We did not use dot pattern stimuli but chose feature patterns that approximated the patterns of rated similarities among dot patterns of different types (Nosofsky & Zaki, 1998, Table 2, p. 251). For the categorization task, we used four studied items, each one-bit distortions (i.e., 211111, 121111, 112111, 111211) of a six element prototype (i.e., 111111). Test patterns consisted of the prototype, a new one-bit distortion (i.e., 111121), and a "random pattern" (i.e., 111122) that differed from each of the studied items by three bits. We used a single stored pattern for the recognition memory task (i.e., 1111111). Two test patterns were used in the recognition simulation: a target (i.e., 111111) and a foil (i.e., 111222), differing from the studied pattern by three bits.

Of note, our use of only one level of distortion in the categorization task during testing followed Nosofsky and Zaki's (1998) inter-item similarity assessments mentioned above. This showed that participants afforded equal ratings when asked to judge the similarity of both low and high distortion test items to previously seen high distortion items presented during the study phase. In contrast to Nosofsky and Zaki's implementation, however, we did not directly use these similarities in our analysis—and the self-similarity of items to themselves was not considered as a free parameter in the model as it was in this earlier work. In our simulation, when a test item was identical to a study item, the two items simply shared all the same features—more generally, we used the set of patterns described above to capture other levels of similarity between study and test items and to produce the approximate equivalent of the rated similarity values reported by Nosofsky and Zaki (1998, Table 2, p. 251).

Model Specifics

The general architecture of the model consisted of recurrence between six featural layers and one conjunctive layer. In the case of the categorization task, the conjunctive layer consisted of four

units corresponding to each of the training exemplars. For the recognition task, the conjunctive layer consisted of a single unit. As in previous applications, the network was considered to have stored studied examples prior to testing. Testing was implemented in the model by presenting external input to the relevant units in the feature layer. A hedged softmax activation function, including the parameter C, was used in the conjunctive layer, and a standard softmax function was used over featural layers in the network. Network temperature was set to 1 throughout this simulation, and the regulatory C parameter was set at a level that produced unbiased responding (i.e., equal proportion of hits and correct rejections). As in Nosofsky and Zaki (1998), this resulted in C parameter settings that differed across the two subject groups and tasks. The probability of endorsing a test item as a category member, or of judging an item to be old in the recognition task, was determined directly from summed overall activity in the conjunctive layer. The difference between the performance of amnesics and control subjects was simulated by a change in the magnitude of connection weights in the network.

Simulation Results

As illustrated in Figure 12, we show that REMERGE is able to reproduce the empirical findings of Knowlton and Squire (1993), based on a difference in weight strength between amnesic and control subjects (i.e., 0.83 vs. 1.50, respectively; see Figure 12 for details of other parameters). As such, the network exhibits relatively preserved categorization performance, but impaired recognition performance, at the lower weight strength (i.e., 0.83), thereby simulating the pattern of findings observed in amnesia.

Figure 13 illustrates how the network's generalization and recognition performance vary as a function of the connection weight strength. It can be seen that generalization rises gradually with weight strength, but recognition performance rises more steeply, thereby explaining how amnesics may perform relatively similarly to control subjects on the categorization task but show a more marked impairment at recognition. It is apparent, therefore, that our model is unable to account for the ability of patient E.P. to perform normally at the categorization task but show chance levels of recognition memory performance (Squire & Knowlton, 1995). While the explanation for this interesting finding remains subject to debate, it has been suggested that successful performance on the categorization, but not the recognition, task may be mediated solely by successful retention in working memory of immediately preceding test items (Palmeri & Flanery, 1999).

Discussion of Findings

We have shown, therefore, that REMERGE is able to reproduce the empirical finding observed by Knowlton and Squire (1993) that patients with amnesia show an impairment in recognition memory despite relatively preserved categorization performance. Interestingly, however, in other settings (e.g., the transitive inference paradigm) we have emphasized that a reduction in the strength of premise pair memories in the network has a relatively greater impact on generalization performance—a profile of performance that contrasts with the greater decrement in recognition memory performance (cf. categorization) observed here.

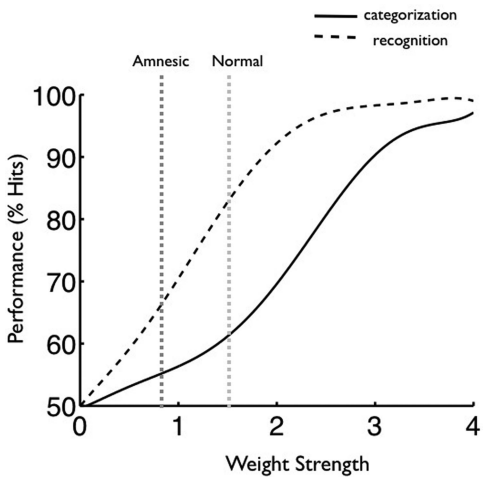

Figure 13. Simulated data relating to the experiment by Knowlton and Squire (1993), illustrating relationship between network weight strength (*x*-axis) and network performance on categorization (solid line) and recognition task (dashed line). Vertical dashed lines indicate weight strength used to simulate the performance of the group of amnesic and normal subjects. See the main text for details.

These simulations, therefore, highlight the different types of generalization involved in tasks such as the transitive inference paradigm (and other inferential tasks) and categorization tasks—with only the former requiring the process of recurrent similarity computation, a process that itself depends on the faithful recall of robustly encoded studied (i.e., premise) experiences. Consistent with this notion, recurrence is not critical to simulating the categorization data obtained in Knowlton and Squire's (1993) experiment—either in GCM, which lacks such a mechanism (Nosofsky & Zaki, 1998), or in an implementation of REMERGE where this function was disabled (data not shown). Instead, successful generalization in this setting may be mediated by the initial weak activation of multiple conjunctive units as a function of their direct, input-based similarity to the test pattern being presented, in a fashion analogous to that performed by classical exemplar models (e.g., GCM).

Taken with the available empirical data (see for reviews: Eichenbaum, 2004; Zeithamova et al., 2012), therefore, we suggest that transitivity, related inferential capacities, and recurrent similarity computation are reliant on the operation of an intact hippocampal system. In contrast, we leave open the much debated question as to whether intact categorization performance—and the underlying process of direct, input-based similarity computation—relies critically upon the hippocampus (e.g., see Squire, Stark, & Clark, 2004; Zaki, 2004), given the lack of decisive empirical data in this regard.

Generalization and Offline Delays/Sleep

Generalized Replay Activity in the Hippocampus Through Recurrency

In the next section, we consider whether REMERGE may also provide a framework for capturing generalization-related phenomena that have been observed in recent sleep studies (Ellenbogen et al., 2007; Gupta et al., 2010). We previously illustrated how

generalization at a behavioral level was driven by graded activation of related elements and conjunctive units in the network. We also introduced the suggestion that graded patterns of activity in the network that support generalization can also be related to neural signals measured by fMRI. Here, we suggest that a recurrent view of this circuit also makes novel predictions concerning the nature, and function, of activity patterns that one would expect to be spontaneously generated during offline periods, such as rest and sleep (e.g., slow-wave sleep [SWS]; Buzsáki, 1989; Diekelmann & Born, 2010; Ego-Stengel & Wilson, 2010; Foster & Wilson, 2006; Girardeau et al., 2009; Gupta et al., 2010; Lee & Wilson, 2002; McClelland et al., 1995; Wilson & McNaughton, 1994).

Recent work has devoted much attention to understanding how patterns of activity in the hippocampus differ between waking and resting periods such as SWS, made possible through the availability of multielectrode recording techniques in rats (reviewed in O'Neill, Pleydell-Bouverie, Dupret, & Csicsvari, 2010). It is known that local field potentials are quite different between waking exploration and resting states, the latter characterized by so-called sharp wave ripple (SWR) complexes, comprised of negative sharp waves and ripple complexes in the CA1 region. Hippocampal ensemble activity during SWS is thought to be shaped by the patterns of synaptic weights formed by experience within the CA3 region and between the CA3 and CA1 regions (O'Neill et al., 2010). In this way, synchronous discharges initiated in CA3, perhaps triggered by neocortical oscillations ("up-states") are viewed to trigger ripples in CA1, whereby waking experience is replayed in time compressed fashion (over 100 ms). Several studies have demonstrated that ripple activity in CA1 reflects the reactivation of recent experiences in linear track environments, expressed as the time compressed sequential firing of location-specific place cells (Foster & Wilson, 2006; Gupta et al., 2010; Lee & Wilson, 2002; Skaggs & McNaughton, 1996). Replay has been shown to occur both in forward (e.g., A–F) and backward (F–A) sequences under such conditions. Ripple activity of this nature is thought to be important for the strengthening of existing memory representations within the hippocampal system itself (Buzsáki, 1989, 1996) and the consolidation of recent information from the hippocampal system to neocortical systems for long term storage (Buzsáki, 1989; Káli & Dayan, 2004; McClelland et al., 1995; O'Neill et al., 2010). In line with the latter notion, ripple activity in CA1 is correlated with activity in neocortical target regions (Ji & Wilson, 2007), and the specific disruption of ripple activity in the CA1 region has been shown to produce an impairment in the consolidation of spatial memories in rats (Ego-Stengel & Wilson, 2010; Girardeau et al., 2009).

A conventional unidirectional view of the hippocampal system suggests that replay activity in CA1 arises due to the reactivation of a CA3 ensemble coding for a single conjunctive experience (e.g., sequence of places visited) and is transmitted via the ERC to neocortical regions as part of the process of consolidation (McClelland et al., 1995; O'Neill et al., 2010; Squire, Cohen, & Nadel, 1984). According to this account, ripple activity in CA1 should reflect the replay of individual episodes, as indeed has often been observed (O'Neill et al., 2010).

Our model, however, which places recurrency as central to an understanding of the hippocampal system, makes very different predictions about the nature and function of replay activity. Specifically, the model provides for the possibility that the hippocampal system may give rise to replay activity through a recurrent activation process involving the interaction of multiple related conjunctive units, rather than simply reflecting the output of single episodes. As such, our model predicts that at least under certain conditions (e.g., low inhibition [i.e., high temperature] states during sleep; cf. Buzsáki, 1989), hippocampal replay should be "generalized," which we believe has important consequences for the putative functions of replay in aiding neocortical learning (see later).

Empirical Findings: Gupta et al. (2010)

A recent experiment by Gupta et al. (2010) provides some support for the hypothesis that replay in CA1 can result from the combination of related episodes, rather than a single episode alone. Gupta et al. used a richer maze environment, rather than a standard linear track environment that may provide little opportunity for useful generalization across related episodes. A two-choice T maze was employed (see Figure 14, upper panel), where rats were trained to run in one direction around the maze. Rewards were provided in the return arm of the maze, for L, R, or alternating turns at the final choice point, depending on the experimental session. Replay activity was observed in CA1, using multielec-

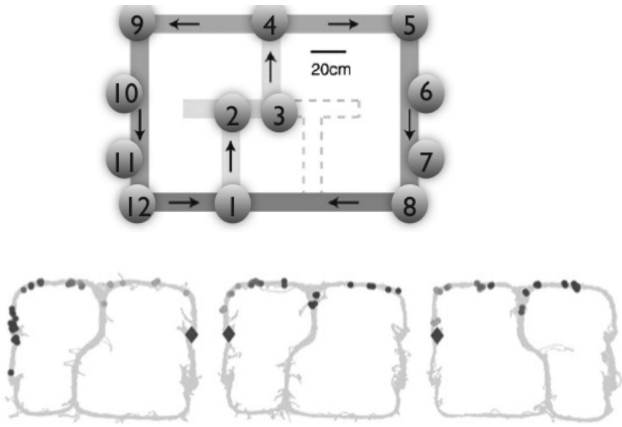

Figure 14. Illustration of replay activity in rodent hippocampus reflecting shortcut sequences (across the top of the maze), and a schematic of maze environment. Adapted with permission from "Hippocampal Replay Is Not a Simple Function of Experience," by A. S. Gupta, M. A. van der Meer, D. S. Touretzky, and A. D. Redish, 2010, *Neuron, 65*, p. 701. Copyright 2010 by Elsevier. Upper panel: schematic of maze environment. Numbers reflect coding of each location—reward was available at Positions 6–7 and 10–11—if rat had made correct turn at Position 4 (which depended on a session specific contingency). In the model, journeys were coded as sets of pairwise paths rather than entire sequences of locations—for example, a left turn journey from Position 1 to Position 12 was coded in the model as individual paths from 1–2, 2–3, and so forth. See the text for details of the model. Bottom panel: points reflect hippocampal spike activity that contributes to sequential replay during ripples (see Gupta et al., 2010, for details). Light gray points reflect spikes that occur earlier in the replay sequence, and dark gray points reflect those that occur later. Three shortcut replay sequences are illustrated, spanning the top of the maze—a journey never actually taken by the rat during waking experience. Gray diamonds indicate location of rat at time of replay activity.

trode recording techniques during periods when rats rested and consumed rewards. As in previous studies, CA1 ripples were observed, reflecting the time compressed replay of sequential place cell activity coding actual paths traversed by the rat during recent and remote waking experience. Critically, however, Gupta et al. observed that never experienced shortcut journeys, for instance a path from top left of the maze to top right, were also replayed within a proportion of SWR complexes (see Figure 14, lower panel). These findings, in conjunction with control analyses demonstrating that the proportion of novel shortcut sequences was significantly greater than would be expected by the chance recombination of individual forward and backward replays, suggest that replay activity in CA1 may reflect the combination of multiple episodes, rather than exclusively individual journeys.

The goal of our simulation was to provide support for the notion that spontaneous activity within a recurrent hippocampal system naturally produces shortcut sequences of the sort observed by Gupta et al. (2010), mediated by the interaction of multiple related conjunctive units.

Details of Model Architecture and Replay Simulation

The generic bidirectionally connected two-layer (feature, conjunctive) model formed the basis for the replay simulations. Units in the feature layer corresponded to significant locations in the maze (i.e., Locations 1–12; see Figure 14), with conjunctive units corresponding to pairwise journeys (e.g., 1–2). As before, weights were fixed, and the network was considered to have stored paths corresponding to paths taken during waking experience. These were coded as pairwise journeys, as positive bidirectional connections between feature layer units (e.g., 1, 2) and a conjunctive unit (1–2). Weights leading from a feature unit corresponding to the start of a journey (e.g., 1) to the relevant conjunctive unit (1–2) were greater (i.e., weight strength = 1) than those leading from the feature unit corresponding to the end (e.g., 2; i.e., weight strength = 0.9). This weight asymmetry resulted in a tendency for the network to replay paths in a forward, rather than reverse direction, consistent with prior empirical data (O'Neill et al., 2010). A logistic sigmoid activation function was used in the feature layer, and a hedged softmax activation function was used over the conjunctive layer. The two free parameters in the network (τ, C) were set at 0.2 and 5, respectively. In order to produce variability in path generation, a small amount of random noise ($M = 0$, $SD = 0.005$) was added to the net input to units.

Replay in our model occurred as follows: (1) The spontaneous activation of pyramidal cells in the hippocampus during SWS states was simulated, as in (Johnson & Redish, 2005), by randomly selecting one unit in the feature layer and providing it with an external input of 1 (e.g., unit 4); (2) the network was allowed to cycle until another unit (except the one receiving external input) crossed a preset threshold (i.e., activation value of 0.6). If more than one unit did so simultaneously, one unit was chosen at random. The winning unit (e.g., 5) was given an external input of 1, and the network allowed to cycle once again. Importantly, the previously active unit (i.e., 4) was prevented from becoming active, in effect being inhibited from participating in the rest of the replay sequence, analogous to an inhibition of return mechanism. This procedure was iterated until a replay sequence of up to three nodes was generated (e.g., 4–5–9). Path lengths of 3 were chosen to mirror the approximate lengths of paths reported by Gupta et al. (2010).

It is important to acknowledge that our model (and previous work; e.g., Johnson & Redish, 2005; also see Molter, Sato, & Yamaguchi, 2007) followed a simplified scheme, designed to mirror the broad characteristics of replay rather than obey specific neurophysiological properties of the hippocampus. As such, we "hard-coded" paths in the model, rather than simulating the actual encoding of paths through the environment, often implemented under theta state conditions facilitating sequence coding in the CA3 region. Further, paths were coded in pairwise, rather than sequential, fashion, both for simplicity and to mirror the coding of premise pairs in previous simulations. Given the inherent limitations of our model, therefore, we aimed to simulate the overall pattern of findings observed by Gupta et al. (2010) rather than fit the detailed quantitative pattern of the data.

Replay sequences were generated using this operating procedure, for a total of 50 instances from each possible starting position in the maze (i.e., unit in the feature layer). As in Gupta et al. (2010), replays were categorized according to whether they reflected individual experiences (forward [e.g., 4–5–6], backward [4–3–2]), shortcut sequences (e.g., 8–4–5), or disjoint sequences (e.g., 4–5–9). The model generated 60% forward replay sequences, 26% backward sequences, 14% shortcut sequences, and $<1\%$ disjoint sequences. We were therefore able to replicate the basic finding of shortcut sequences observed by Gupta et al. (2010), though in a greater proportion than observed empirically. This difference may have arisen for a variety of different reasons (e.g., strength of weights in the network): As noted above, our aim was to confirm the basic capacity of the network to generate novel shortcut sequences during replay, rather than achieve a close fit to the empirical data through optimizing the parameters of the model.

Summary of Model Simulations and Conclusions

1. Spontaneous activity in the hippocampus during offline periods is generally thought to reflect the reactivation of a single CA3 ensemble coding for an individual episode. Here, we verify the basic capacity of a recurrent network to produce what we call "generalized" replay.

2. We describe a simulation of the model that produces patterns of replay activity reflecting never-experienced shortcut paths, replicating the basic finding observed in a recent empirical study by Gupta et al. (2010).

3. In previous simulations, we illustrated how recurrency, through the co-activation of multiple conjunctive units for related episodes, could support generalization at a behavioral level (e.g., in the transitive inference task). It is worth noting that while the set of network parameters (e.g., temperature) employed in the current replay simulations resulted in the co-activation of multiple conjunctive units, generalized replay could in principle be mediated by a "chaining" process involving the sequential activation of individual conjunctive units.

In addition to simplifications alluded to above (e.g., coding of sequential journeys in pairwise fashion in the model), we also wish to point out that our model was only designed to simulate shortcut paths that reflect combinations of previously taken journeys—as observed by Gupta et al. (2010). In contrast, the ability of rats to navigate to a particular location (e.g., hidden platform in a water

maze; e.g., Eichenbaum, Stewart, & Morris, 1990) from a range of novel starting points falls outside the scope of our simulations, perhaps reflecting the use of an allocentric map of space based on specifically geometric computations (O'Keefe & Nadel, 1978). In a later section (see General Discussion), we outline empirical predictions that derive from our notion of generalized replay activity in the hippocampal system—and we discuss how this may have broad implications for the function of the hippocampal–neocortical dialogue.

The Emergence of a Capacity for Transitive Inference Over an Offline Delay: Ellenbogen et al. (2007)

We next examined the intriguing finding that transitivity emerged over an offline delay during which subjects rested or slept, without receiving further additional training (Ellenbogen et al., 2007). We first consider whether this striking enhancement of inferential performance over an offline delay might also be accounted for by the strengthening of premise pair memories—as we previously suggested could mediate the effect of continued training in a previously outlined simulation of the transitive inference task (i.e., in Ryan et al., 2009; see Figure 5).

Fifty-six healthy volunteers participated in the experiment by Ellenbogen et al. (2007). There were two principal subject groups of interest, defined by the offline period (20 min, 12 hr) that intervened between premise pair training (e.g., A–B) and testing on inference pairs (e.g., B–D, B–E, etc.). The stimuli and design of the study were similar to Ryan et al. (2009), though a different training regime was used. During training, premise pairs were presented in pseudorandom order (e.g., B–C could not follow A–B), a manipulation designed to avoid revealing the structure of the hierarchy to subjects. Premise pair training continued until subjects reached a predefined criterion, which was set at greater than 75% on the middle premise pairs (i.e., B–C, etc.). Subjects then completed an immediate test on the premise pairs, without feedback, followed by a variable offline delay (e.g., 20 min or 12 hr). Following this delay, subjects were tested with inference pairs and premise pairs, once again without feedback. Subjects were asked to report their confidence in their choices during inference trials, though explicit awareness of the hierarchy was not assessed in this study.

As shown in Figure 15 (upper panel), despite successful performance on premise pairs (e.g., B–C) after the initial training session (c90%), inferential performance was not significantly different from chance levels, shortly after the completion of training (i.e., following a 20-min offline delay). Strikingly, a capacity for transitivity emerged after an offline delay of 12 hr (or more): Inferential performance, averaged across both "close" (i.e., B–D, C–E) and "distant" (i.e., B–E) inference pairs, was 75%.

The results of our simulation of this experiment are shown in Figure 15 (lower panel). As such, it can be seen that the network is able to simulate the empirical finding that a capacity for inference appears only after an offline delay with little change in premise performance. Notably, the network is able to reproduce this phenomenon by appealing to a simple mechanism by which the strength of premise pair memories increases over the offline delay—analogous to the mechanism previously proposed to underlie the effects of continued training. As discussed previously, the large lag generalization effect observed in this simulation

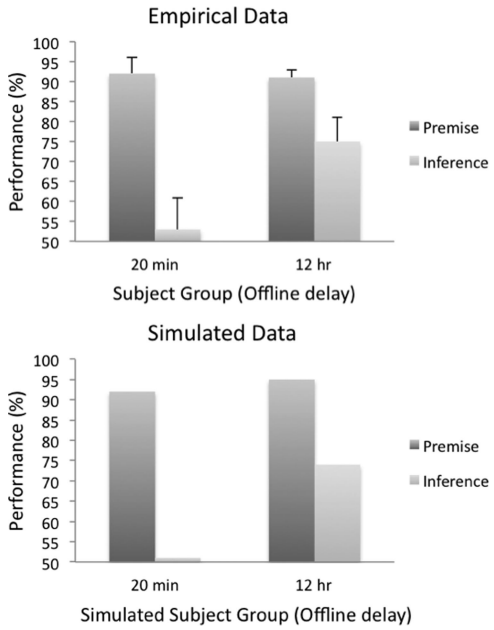

Figure 15. Transitive inference task: empirical (upper panel) and simulated (lower panel) data from Ellenbogen et al.'s (2007) study. Performance of 20-min and 12-hr subject groups shown (*x*-axis) with performance (%) on *y*-axis. Groups differ as a function of the length and nature of the delay period interposed between training and testing. Premise performance (dark gray bar) averaged across all relevant pairs (i.e., A–B, B–C, . . . E–F). Inference performance (light gray bar) averaged over close (B–D, C–E) and distant inference (B–E) pairs. Parameters: $\tau = 0.2$; $C = 15$; $\beta = 0.3$. Weight strengths for the 20-min and 12-hr groups, respectively: 1.29, 1.35. Error bars (empirical data) reflect standard error. Model performance (expressed as percentage correct) derived directly from Luce choice ratios (see the main text for details).

arises as a consequence of the relatively low network temperature (i.e., 0.2; see Figure 4, lower panel).

While the network is able to simulate the sudden emergence of inferential capacity over an offline delay, two points suggest that such a weight strengthening account falls short of providing a complete account of such generalization-related sleep phenomena. First, the network fails to reproduce the more specific benefit of sleep on inferential capacity reported by Ellenbogen et al. (2007), whereby performance on distant pairs (i.e., B–E) was seen to improve to a greater extent compared to performance on close pairs (i.e., B–D, C–E; empirical data not shown). Indeed our model, implementing recurrent similarity based generalization, is unable to provide a robust account for this phenomenon (i.e., B–E > B–D). Specifically, performance on close inference pairs in our model tends to emerge at lower, or similar, weight strengths, compared to that of distant inference pairs, depending on the exact parameter settings (i.e., temperature, beta: results not shown).

Second, it has often been argued that offline periods result in mnemonic benefits (e.g., in paired associate paradigms) through a reduction in forgetting (Diekelmann & Born, 2010; Wixted, 2004). Indeed, it is typically the case that memory performance, for instance on a paired associate recall task, is actually worse after a period of sleep compared to immediately after training, though significantly superior to a comparable period of waking activity

(though see Tucker & Fishbein, 2008, for an example of better memory after sleep compared to immediately post-training). Notably, the simulation above relies on the idea that sleep leads to the strengthening of recent memories. While recent work suggests that performance on instrumental tasks involving rewards may actually be enhanced as a result of sleep—a situation that has some similarities to feedback-based learning tasks such as the transitive inference paradigm—it remains uncertain whether such an effect would generalize to the kind of experimental scenario considered here (Brawn, Nusbaum, & Margoliash, 2010).

How then might the beneficial effect of an offline delay on the emergence of a capacity for inference be accounted for? One hypothesis, favored by Ellenbogen et al. (2007), is that offline delays may facilitate the creation of stored generalizations (e.g., coding the linear structure of the transitivity task) through the recombination of multiple related episodes, which can then be used as the basis of performance on inference trials. In the previous section, we presented empirical evidence suggesting that replay activity during offline periods may indeed reflect the recombination of related episodes and therefore be generalized in nature, and we demonstrated that REMERGE is able to reproduce this basic phenomenon. While we do not formally explore the putative creation of stored generalizations in the current article, given our use of an idealized learning scheme and the potential pitfalls of allowing new learning to coexist with memory retrieval (e.g., Hasselmo, 1993), we believe the recurrent architecture we advocate constitutes a natural framework in which to consider these ideas (see General Discussion).

General Discussion

Overall Summary

Empirical evidence supports an important function for the hippocampus in generalization, in addition to its well-established role in episodic memory (Eichenbaum, 2004; Zeithamova et al., 2012). Here, we ask how the hippocampus mediates generalization, given that this brain structure, and in particular the DG and CA3 subregions, is typically viewed to perform pattern separation (i.e., orthogonalization) of incoming inputs—a process that emphasizes the differences between related experiences at the cost of disregarding their shared features. Our approach to this issue involves "big-loop" recurrency within the hippocampal system, or more widely between the hippocampal system and the neocortex, which mediates the interaction of pattern separated conjunctive codes (broadly ascribed to the CA3 region) with componential representations viewed to be instantiated in regions of the neocortex. As such, generalization in the REMERGE model occurs at the stage of memory retrieval and results from recurrent similarity computation, with the activity of conjunctive units reflecting their similarity to both externally presented sensory inputs, and inputs reconstructed by the network. Recurrence, therefore, expands the scope of exemplar and similarity based models of generalization to include inferential paradigms such as the transitivity task, while preserving the abilities of these models to perform categorization and recognition memory.

We also offer a mechanistic account of how a capacity for generalization may emerge through training (e.g., in Ryan et al., 2009)—or differ between subjects (e.g., Shohamy & Wagner, 2008)—simply through the strengthening of existing memories for premise items, rather than necessitating the creation of explicit representations of the task structure. Our theory, therefore, offers a formal model of how hippocampal generalization emerges, as if within a virtual memory space as previously hypothesized (Eichenbaum et al., 1999), but where the linkage of related experiences occurs on the fly through recurrent interactions, rather than being statically engrained in the actual overlap of neuronal codes at the representational level. As such, the current perspective may help reconcile fundamental differences between two highly influential views of memory: the CLS (McClelland et al., 1995) and relational theory (Cohen & Eichenbaum, 1993; Eichenbaum, 2004).

In this section, we discuss the relationship between our retrieval-based theory of generalization and encoding-based overlap accounts that emphasize the importance of creating overlapping representations for related experiences during training (i.e., encoding). We first consider the relationship between our account and a formal encoding-based overlap theory: the temporal context model (TCM; Howard et al., 2005). We then turn to points of contact between our account and the relational theory of hippocampal function (Cohen & Eichenbaum, 1993; Eichenbaum, 1999), before examining how our theory fits with the empirical observation that hippocampal neurons have been observed to respond to common features across multiple experiences (i.e., "nodal codings"). Next, we outline predictions that arise from our notion of generalized replay in the hippocampal system. Finally, we discuss implications for the individual contributions of the hippocampus and neocortex to semantic learning, construed as learning the general structure of the environment.

REMERGE and TCM: Key Differences and Divergent Predictions

We first examine the relationship between REMERGE and the TCM, a formal mechanistic account that proposes that it is the overlap between hippocampal representations for related stimuli (or experiences) that is critical to generalization (Howard et al., 2005). TCM was originally developed to account for essential properties of behavioral data on tasks involving free recall (Kahana, 1996; Polyn & Kahana, 2008; Polyn, Norman, & Kahana, 2009) but has more recently been proposed to account for inferential behavior in the transitive inference task (Howard et al., 2005). Briefly, TCM argues that contextual states, rather than item–item associations, act as the primary cues for recall of items. As such, items are retrieved as a function of their similarity to the current state of context, which in turn is influenced by both the items themselves and a general tendency to drift over time.

To summarize the essential principle of the mechanism of generalization by TCM and to illustrate how this differs from that of REMERGE, we use the transitive inference paradigm as an illustrative scenario. During training, the contextual representation of each item in a premise pair (e.g., Items A and B) shares common features due to their temporal co-occurrence. Critically, the item-contextual codes of non-adjacent items—for example, A and C—also overlap: The reason for this is that during BC training, the C item is bound to the current state of context, which includes the retrieval of the context in which B was encoded (i.e., co-occurrence with A). Over training, therefore, the model develops a

similarity structure that comes to reflect the distance of items within the linear hierarchy of the transitive inference task. While to our knowledge, simulations of the model's behavioral performance on the transitive inference paradigm have not been presented, it is argued that successful inference could be mediated by a value gradient (i.e., B > E) constructed from TCM's linear representation of items. Further, it is proposed that the hippocampus is critical to transitivity judgments through its capacity to support new item-contextual learning.

TCM, therefore, exhibits the phenomenon of transitivity in a qualitatively different way from REMERGE. While inference in REMERGE can be considered an emergent phenomenon at retrieval mediated by recurrency operating over pattern separated conjunctive codes, the same inference in TCM is dependent on the creation at encoding of overlapping item-contextual codes whose similarity to one another reflects the relationship of items in a set of related experiences. Here, we focus on this key difference between REMERGE and TCM that parallels a distinction drawn between model-based and model-free controllers influencing behavioral choice in situations involving reward learning (e.g., Daw, Niv, & Dayan, 2005). In a later section, we consider possible differences between REMERGE and TCM in terms of predicted content of hippocampal replay activity during offline periods.

Model-free systems typically store cached values acquired through learning, for example, concerning the value of stimuli or actions divorced from their respective outcomes (Daw et al., 2005). In contrast, model-based systems construct predictions on the fly, effectively at the point of choice, through a retrieval-based process of search involving the chaining together of short-term predictions concerning the outcome of individual actions in a sequence. While model-free systems are computationally efficient, they are relatively inflexible and tend to resist the updating of their representations in the face of important changes in the structure of the environment (cf. model-based systems). The REMERGE network bears similarities to a model-based system, given that generalization emerges through a retrieval-based search-like process over the space of conjunctive experiences mediated by recurrency. TCM, in contrast, can be construed in broad terms as a model-free system, though it should be noted that TCM learns item representations that are useful for generalization, rather than cached values. Expressing the difference between REMERGE and TCM in these terms may be useful in suggesting possible empirical avenues for distinguishing between them, a possibility we consider next.

REMERGE and TCM make different predictions concerning the effects of changing the structure of the environment on subjects' behavior in the transitive inference and other related tasks. While it may be difficult to definitely distinguish between these competing models of generalization on the basis of any single experiment, it is potentially informative to consider how subjects would respond in the following scenario: The initial phases of the experiment would be identical to a typical five-pair transitive inference task. Next, subjects would receive training on a new premise pair, where F should be chosen over A (i.e., effectively changing the structure of the set of premise pairs from a linear hierarchy to a circle). Under these conditions, REMERGE makes the prediction, which we verify in an additional simulation (see Figure 16), that subjects would be equally likely to choose B over E in a subsequent test trial, given its retrieval-based similarity mechanism of generalization. In contrast, TCM, along with value transfer mechanisms (Frank et al., 2003; von Fersen et al., 1991), would suggest that subjects should continue to favor item B. In TCM, this would be the case because learning of a novel premise pair (i.e., F+ A−) would not generally be expected to change previously encoded item-contextual representations associated with other items (e.g., of B). Of note, both REMERGE and TCM would predict that subjects would continue to favor B over D in a B–D trial, albeit through different mechanisms. While preliminary investigations in chimpanzees (Gillan, 1981) and rats (Davis, 1992) involving small numbers of experimental subjects suggest that transitive responding may indeed be disrupted by the addition of a new inconsistent premise pair (e.g., F+ A−), more extensive testing would be potentially illuminating in distinguishing between these two models of transitivity behavior.

REMERGE and the Relational Theory

Our perspective helps reconcile key differences between two highly influential views of memory—the CLS (McClelland et al., 1995) and relational theories (Cohen & Eichenbaum, 1993; Eichenbaum, 2004)—that to our knowledge have rarely been explicitly acknowledged in the literature to date. CLS theory, in its original form, prioritized the role of the hippocampus in episodic memory instantiated in a pattern separated representation scheme, at the expense of generalization capacity. In contrast, the relational theory (Cohen & Eichenbaum, 1993) has long regarded generalization and semantic learning to be a fundamental aspect of hippocampal function, without such a clear focus on the nature of computational constraints that support optimal performance (e.g., capacity) in a neural system widely agreed to support episodic memory. Indeed, recent formulations of the relational theory, drawing on the empirical phenomenon of "nodal codings" (see below; also see Eichenbaum et al., 1999), have argued that the networking of related episodic memories—based on their overlap at the representational level—is central to hippocampal processing and critical for generalization (Eichenbaum et al., 1999).

The present account of the hippocampal system, instantiated in the REMERGE model, can be considered to marry essential insights provided by a relational view of memory (i.e., compositionality of episodic memories; generalization through linking of related episodes within a memory space; Cohen & Eichenbaum, 1993; Eichenbaum et al., 1999; also see Buzsáki, 2005) with the computational constraints of an episodic memory system imposed by CLS theory and allied viewpoints (i.e., rapid learning, pattern separated representations; Burgess, 2006; Marr, 1971; McClelland et al., 1995; McNaughton & Morris, 1987; O'Reilly & Rudy, 2001; Rolls & Kesner, 2006). This synthesis arises through exploiting the principle of recurrency, which affords the network a capacity to generalize efficiently but critically still preserves the essential characteristics of pattern separated codes in the hippocampus. The account we offer, therefore, seeks to preserve the foundational division between representational schemes employed by the hippocampus (pattern separated) and neocortex (overlapping codes for related experiences; Marr, 1971; McClelland et al., 1995) but considerably expands the potential contribution of the former to generalization, by incorporating recurrent flow within the circuit. In this way, we have been able to capture empirical data implicating the hippocampus in tasks involving generalization (Eichenbaum, 2004; Zeithamova et al., 2012), which until now

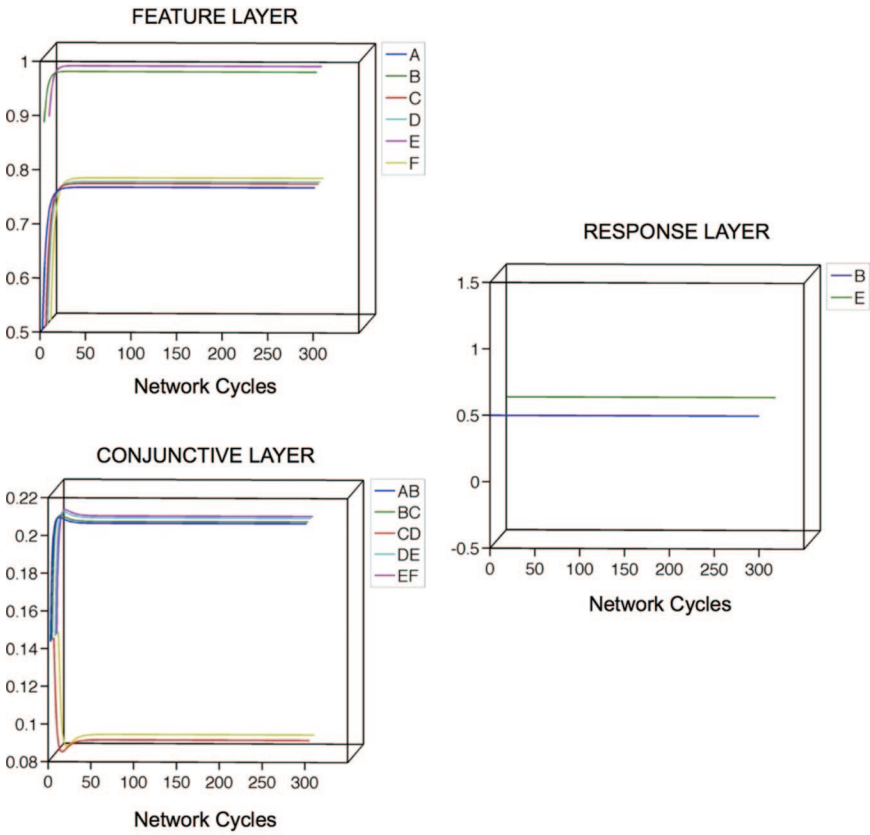

Figure 16. Transitive inference task: Illustration showing that the network no longer favors B over E in a BE trial, following the addition of the F+ A− premise pair to the existing set of premise pairs—which effectively transforms the linear hierarchical arrangement of stimuli into a circular configuration (see the main text for details). Parameters: $\tau = 0.25$; $C = 1$. Weight strength = 1 (i.e., as in simulation shown in Figure 2).

have tended to speak exclusively to the relational theory, residing as they have done largely outside the explanatory framework of CLS theory.

Pattern Separation Versus Nodal Codings?

Our approach emphasizes the importance of considering how the hippocampus might contribute to generalization based on pattern separated neural codes for conjunctive experiences. This contrasts with encoding-based overlap accounts, including the TCM and relational theory (Eichenbaum, 1999; Gluck & Myers, 1993; Howard et al., 2005; also see O'Reilly & Rudy, 2001). These accounts embrace the notion that it is the overlap between representations of similar episodes (or stimuli) that is critical to inference (e.g., in the transitivity task). We next consider empirical evidence that has been interpreted as supporting this argument, and we ask whether such observations are consistent with our perspective.

In one important study (Wood, Dudchenko, & Eichenbaum, 1999; also see Eichenbaum, 1999; Ferbinteanu & Shapiro, 2003; Moita, Rosis, Zhou, LeDoux, & Blair, 2003), activity of neurons in the hippocampus was recorded while rats performed a continuous non-match-to-sample task. They observed hippocampal neurons with a range of specificities, varying from some, termed *nodal codings* that responded to a common feature across many events (e.g., a particular odor), to others that showed more selective conjunctive responses (e.g., to a non-match decision performed in relation to a particular odor at a particular place). How then can the presence of hippocampal nodal cells encoding common features be accommodated by our perspective and related perspectives (e.g., Marr, 1971; McNaughton & Morris, 1987; Treves & Rolls, 1992), espousing an important role for the hippocampus in pattern separation?

Two points are worth highlighting in the current context: First, in the current perspective, we adopt a theoretical ideal of pattern separation (i.e., zero overlap of conjunctive codes) to illustrate how generalization can still be achieved through recurrency under these circumstances. We naturally accept, however, that neural representations of similar episodes (or environments) are likely to overlap since of course in reality the process of pattern separation is far from perfect. Such "residual" overlap of neural codes within the CA3 region might be viewed to arise from a failure of optimal pattern separation of incoming inputs that already overlap at the level of upstream regions (i.e., ERC), rather than a specific hippocampal mechanism designed to support generalization per se. Nevertheless, it is worth bearing in mind that residual overlap of this nature may support a degree of "short-range" generalization,

even in the absence of recurrency. For example, it has been shown that successful B–D, but not B–E, test trial performance in the transitivity task may be mediated by such a mechanism (O'Reilly and Rudy, 2001).

Second, it should be emphasized that pattern separation is most appropriately thought of as a function of part of the hippocampal circuit, rather than the system as a whole (McNaughton & Morris, 1987; O'Reilly & McClelland, 1994). As discussed previously, it is generally agreed that information processing within the DG and CA3 regions is particularly important for creating and sustaining orthogonalized codes for similar input patterns. This view is supported by empirical data suggesting that neural codes for similar environments in DG and CA3 overlap less than those in CA1. In a recent experiment (S. Leutgeb et al., 2004), neural activity was recorded in both CA3 and CA1 while rats foraged in different enclosures (e.g., square, circle) in one of three different rooms (A, B, C). Importantly, they noted a striking difference between CA3 and CA1 in terms of the neural representations observed in each of the different room configurations. While the pattern of neural activity in CA3 was essentially orthogonal in each environmental setting, CA1 responded to common features, analogous to the nodal codings reported by Wood et al. (1999): For example, the pattern of CA1 activity in the square enclosure was highly correlated (i.e., similar) regardless of the room in which it was situated.

This evidence suggests that neural codes in DG-CA3 may be viewed as more conjunctive and pattern separated, compared to those in CA1 that are correspondingly more overlapping and componential. Related evidence using similar experimental settings is also consistent with this conclusion (J. K. Leutgeb et al., 2007; Skaggs & McNaughton, 1998; Tanila, 1999). While it is not clear if the nodal codings (e.g., odor cells) observed by Wood et al. (1999) tended to be situated in the CA1 region, with more specific codings in CA3 (e.g., decision-odor-place codings), such a pattern of findings would be consistent with this overall scheme.

It is also worth asking how hippocampal neurons (e.g., in CA1) that code for common features of an environment might arise, and what their functional significance might be. One hypothesis is that nodal codings are driven by feedforward connections from the ERC to CA1, as proposed by McClelland and Goddard (1996). While further empirical work is necessary, such a mechanism has been suggested to be consistent with the rapid timecourse (cf. CA3) of the development of CA1 representations in S. Leutgeb et al.'s (2004) study. If this were the case, this might suggest that nodal codings in the hippocampus do not necessarily have any special functional role in generalization (i.e., as suggested by the relational theory). Instead, they might reflect the use of a coding scheme in CA1 that is relatively more componential than that sustained by CA3 to allow information arriving from CA3 (e.g., as hippocampal output during waking, or during replay states) to be transformed into a format that can be interpreted by other brain regions (McClelland & Goddard, 1996).

Our suggestion, therefore, is that nodal codings in the rodent hippocampus may reflect the componential nature of inputs (and outputs) of the DG–CA3 regions, rather than being indicative of a specific hippocampal mechanism for generalization per se. As noted previously, we believe that structured inputs/outputs of this nature are critical to the rapid creation of episodic representations comprised of arbitrary conjunctions of multiple components (e.g., see McClelland & Goddard, 1996)—with generalization then relying on a recurrent mechanism that operates over the set of conjunctive codes. In accordance with this general position, we broadly identify the conjunctive layer of our model to the type of codes sustained by the DG–CA3 regions, and the feature layer to the CA1 and upstream neocortical regions including the ERC. We note, however, that it is possible that neurons sustaining conjunctive codes may interdigitate with those supporting relatively more componential codes, rather than being confined to anatomically distinct subregions.

At this stage, it is also worth drawing attention to a type of hippocampal neuronal response profile that bears some similarities to the nodal codings described above. Neurons in the hippocampi of epilepsy patients have been found that appear to respond to stimuli that denote a particular individual (e.g., the written name or a picture relating to Jennifer Aniston) and associated entities (e.g., Brad Pitt), rather than being specific for any unique episodic experience (Quiroga et al., 2008; Quiroga, Reddy, Kreiman, Koch, & Fried, 2005). Interestingly, however, neuronal response selectivities of this sort have typically been reported to occur throughout the medial temporal lobe (including the amygdala) and at very long latencies (e.g., 300–600 ms; Quiroga et al., 2008, 2005). While these observations may also be accounted for along the lines discussed previously (i.e., as part of a relatively componential code sustained by regions upstream of CA3, or even interdigitated with more conjunctive neurons in CA3), one additional possibility suggested by their surprisingly long response latencies is that such neurons are not primarily driven by a feedforward mechanism but reflect the operation of a recurrent mechanism involving the coactivation of multiple related conjunctive ensembles. While we note, therefore, that the set of observed findings tends to be consistent with the view that neurons throughout the MTL may respond to abstract entities such as individuals or objects that are present in many different episodes (Quiroga et al., 2008), the presence of such neurons may not be inconsistent with the possibility that generalizations across experiences are formed by recurrent computations involving more conjunctive representations.

Comment on Stored Generalizations

The notion that stored generalizations, often termed integrated representations, are critical to generalization is common to several accounts of inferential behavior in tasks such as the transitive inference, paired associate inference, and acquired equivalence paradigms (Eichenbaum, 2004; Howard et al., 2005; Shohamy & Wagner, 2008; Zeithamova & Preston, 2010). According to the relational theory discussed previously (Cohen & Eichenbaum, 1993; Eichenbaum et al., 1999), it is the overlap in neural representations between related episodes that supports discovery and, presumably, explicit representation of the task structure (e.g., A–B–C. . .–F: linear hierarchy in the transitive inference task). The integrative encoding hypothesis, in contrast, suggests that it is the pattern completion properties of the hippocampus mediating the retrieval and subsequent encoding of related episodes into a larger representation of the task structure (e.g., linking stimuli F1, F2, S1, S2 in their experiment; e.g., see above), that are critical to inference (Shohamy & Wagner, 2008).

The REMERGE model places particular emphasis on the idea that a capacity for generalization is an emergent property of the recurrent network we describe, effectively acting at the stage of memory

retrieval. As such, determining when such stored generalizations, or integrated representations, formed at encoding are needed to account for behavioral performance represents an important challenge for future investigators. That said, we do acknowledge that stored generalizations may well arise during training, testing, and intervening periods including but not limited to sleep, and that they may thereby make a contribution to generalization and inferential behavior. For instance, it seems clear that the capacity to explicitly report details of the task structure (e.g., the linear hierarchy in a transitive inference task) can emerge in human subjects under certain experimental conditions (Moses et al., 2006; C. Smith & Squire, 2005; also see Kumaran et al., 2007, 2009). Such explicit reports and other behavioral indices of generalization could reflect processes that occur as a consequence of recurrent similarity-based generalization as captured in the REMERGE model.

Therefore, we would not wish to exclude the possibility that the hippocampus creates new representations combining elements from distinct experiences that became co-activated through recurrent processing. Indeed, the notion that emergent generalizations once created through a REMERGE-like process would then themselves be stored in memory is very appealing. It will remain for future research to evaluate how easily the storage of such representations could co-exist with the episodic memory function of the hippocampus and related areas. More generally, the potential perils of allowing new learning to be intertwined with network recall should not be underestimated (Hasselmo & Bower, 1992; Hasselmo & Schnell, 1994).

These caveats not withstanding, if weights were allowed to change during network operation during learning or offline periods such as slow-wave sleep (see below), it is reasonable to imagine that, under certain circumstances (i.e., depending on learning algorithm and inhibitory control), a new conjunctive representation capturing the graded patterns of activity present on the feature layer would result in stored generalizations of the sort envisioned by Shohamy and Wagner (2008). Importantly, stored generalizations, created in this manner, would be coded by distinct neural representations that would be simulated using their own localist conjunctive units in the model, separate from those pertaining to the individual premise pairs, thereby limiting the potential detrimental effect on the network's ability to recall specific experiences (i.e., AB).

Generalized Replay in the Hippocampal System: Perspective and Predictions

In line with the ideas considered above, we suggest that generalized replay activity in the hippocampal system may facilitate the creation of new representations (i.e., stored generalizations), formed from the recombination of multiple related episodes—as well as the strengthening of existing hippocampal representations (i.e., intrahippocampal consolidation; Buzsáki, 2005). As such, we draw on previous ideas that replay activity may be an important determinant of plasticity within the hippocampal system itself, by bringing episodes spanning relatively long timescales (e.g., tens of seconds) within a time window appropriate for linking them through associative synaptic modification (e.g., 100 ms; Buzsáki, 1989, 1996, 2005). Interestingly, this general idea has been proposed to underlie the transition from direction-selective, to omnidirectional, place cell responses in the hippocampus (Buzsáki, 2005).

We also suggest the following predictions concerning generalized replay in the hippocampal system:

1. Our model predicts that generalized replay should be observed in a wide range of settings, for example, an offline period following a block of transitive inference task performance. Further, a selective impairment of the potential for recurrency within the hippocampal system would be predicted to abolish generalized replay activity. While existing neuroanatomical evidence is consistent with a potential for recurrency in the hippocampal system (van Strien et al., 2009), further work is required to define more precisely the architecture of the relevant neural circuitry to assess the feasibility of such an experimental manipulation. Of note, one possible explanation for the infrequent detection of generalized replay thus far is the relatively simple structure of rodent tasks in which hippocampal replay activity is typically recorded, which tend to offer little opportunity for recurrent similarity-based generalization given their inherent lack of overlapping experiences in the environment (though see Gupta et al., 2010).

2. In the model, graded patterns of activity in the network emerge with the strengthening of premise pair memories during continued training and underlie both behavioral generalization performance and generalized replay during offline periods. Consequently, generalized replay may serve as a marker of the capacity of the system to support behavioral generalization at a given stage of training. A resulting prediction, therefore, is that replay activity during offline periods should arise in parallel with a capacity for behavioral generalization during waking experience.

3. As noted previously, hippocampal replay activity is thought to play a role in neocortical (McClelland et al., 1995; O'Neill et al., 2010) as well as intrahippocampal learning—both in strengthening existing memory representations for individual episodes (Buzsáki, 1989, 1996) and, we suggest, in facilitating the creation of novel stored generalizations. Our model makes the prediction that generalization performance should be relatively susceptible to the disruption of replay-mediated intrahippocampal learning processes, given the particular sensitivity of this behavioral measure to the strength of premise pair representations in the network as well as the possible influence of stored generalizations on behavioral performance. While recent work has demonstrated the feasibility of detecting ripple activity within the hippocampus and has examined the consequences of its disruption on (spatial) memory (Ego-Stengel & Wilson, 2010; Girardeau et al., 2009), the specific consequences for generalization performance have not yet been explored.

The putative role of generalized replay in facilitating intrahippocampal learning also suggests that subjective reports pertaining to neural activity during sleep may correlate with subsequent generalization performance. While such "dream reports" have been typically linked to REM sleep, they have also been associated with SWS sleep, a phase in which hippocampal replay activity is most prominent (Wamsley, Tucker, Payne, Benavides, & Stickgold, 2010). Our model predicts that the frequency of dream imagery reflecting the recombination of multiple related episodes experienced during waking (e.g., a shortcut through a maze) may correlate with subsequent generalization performance (e.g., shortcuts, transitivity judgments). One recent study reports preliminary evidence consistent with this prediction: The frequency with which subjects reported SWS associated dreaming about a maze like

environment in which they had been previously trained showed a significant correlation with subsequent behavioral performance (Wamsley et al., 2010). Interestingly, dream reports in this study tended to be fragmentary, and did not relate to any single journey through the maze. However, they were rather non-specific, precluding any specific conclusions about their relationship to the content of neural replay activity.

Our account links the content of replay activity produced by a recurrent hippocampal system during SWS to a capacity for generalization. Interestingly, however, previous work has suggested that REM sleep, rather than SWS, facilitates generalization (Walker & Stickgold, 2010). Importantly, however, the benefits of REM sleep in terms of generalization have been observed in tasks like the Semantic Priming Test (Stickgold, Scott, Rittenhouse, & Hobson, 1999) and the Remote Association Test (Cai, Mednick, Harrison, Kanady, & Mednick, 2009)—which we consider to be qualitatively different from the transitive inference (and related) tasks and arguably not dependent on the hippocampus (though no data speak to this issue, to our knowledge). In our view, the critical feature of the transitivity task that accounts for its dependence on the hippocampus is the arbitrary nature of the associations learnt (i.e., premise pairs), with inference supported by a capacity to relate multiple premise pairs to one another. This is not the case in tasks such as the remote association test—in the study by Cai et al. (2009), for example, a period of REM sleep was reported to increase the extent to which participants could be primed to identify the word that links a triplet of distantly related associates (e.g., associates: sixteen, heart, candy/linking word = sweet). In this experimental context, REM sleep mediated modifications (e.g., in terms of connectivity) to existing neocortical representations for the familiar stimuli used (e.g., "sixteen" and "sweet") may be sufficient to mediate performance enhancements (see Cai et al., 2009, for a related argument) without involving the hippocampus.

It is worth noting that representational overlap theories (e.g., TCM) might also make the prediction that replay activity in the hippocampal system during offline periods should be generalized in nature, rather than relating solely to a single experience, and should correlate with behavioral indices of generalization. While to our knowledge this issue has not been considered previously, it is therefore conceivable that overlap-based accounts might make many of the same predictions concerning the function of generalized replay in the hippocampal system. There is a possible divergence of predictions, however: REMERGE predicts that periods of generalized replay may be interspersed between periods of conventional single-episode replay (e.g., AB episode), given that network output is determined by the operating level of inhibition that is likely to vary over the course of an offline delay (cf. Buzsáki, 1989). Overlap-based theories (e.g., TCM) would find it difficult to account for such a pattern of replay activity—this is because the similarity between related episodes (i.e., AB and BC) is viewed to be statically engrained at the representational level, rather than dynamically computed at the point of retrieval as in REMERGE.

Perspective on the Role of the Neocortex in Generalization and Semantic Learning

The neocortex is widely held to be a powerful learner of the general properties of the environment but is thought to have several important limitations (McClelland et al., 1995; Rogers & McClelland, 2004). Learning in neocortical networks, at least for information that is unrelated to or inconsistent with existing knowledge, must necessarily proceed slowly, generally requiring many hundreds or thousands of training trials. This is the case because rapid learning of such information can lead to catastrophic forgetting of previously acquired information (McCloskey & Cohen, 1989). Learning in neocortical networks is also constrained by a requirement for interleaved training, whereby training examples inconsistent with pre-existing structure or encoding a new set of structural relationships must be intermixed with familiar information, to ensure optimal extraction of the structure of the data set and integration with pre-existing knowledge without producing catastrophic interference. Focused learning (repeated presentation of the same new items without interleaving) tends to have the same effect as use of large connection weight changes, also tending to produce catastrophic interference when the new items are inconsistent with what is already known (Grossberg, 1987; McCloskey & Cohen, 1989; see also French, 1999).

The principle of slow learning in the neocortex is particularly problematic given the paucity of learning opportunities in the real world: Consequently, neocortical learning is typically viewed to require help from a complementary learning system, putatively instantiated in the hippocampus, capable of serving as the basis of initial learning and providing additional training trials during offline periods—and additionally acting as a basis for memory for recent episodes (McClelland et al., 1995). Although these provisions certainly help to overcome some of the shortcomings of neocortical learning, several open questions remain, which we suggest might be addressed by recurrency in the hippocampal system.

Proposal of How a Recurrent Hippocampal System May Mitigate Limitations of Neocortical Semantic Learning

1. If neocortical learning is very gradual, how are individuals able to rapidly exploit the relationships between a set of related conjunctive experiences, for example, within a single experimental session? We argue that a recurrent mechanism may underpin the contribution of the hippocampus to generalization in the context of the transitivity and related paradigms. Interestingly, very few studies to date have investigated conceptual learning and generalization in a more naturalistic context (e.g., learning and generalizing about a fictitious set of characters as simulated in McClelland, 1981). We suggest that the hippocampal system may make a particular contribution to generalization in settings where the environment is volatile or experience is limited (i.e., few training exposures), whereas the neocortex may prevail in knowledge representation for information acquired over more extended time periods. The greater the temporal extent, the greater the likely role of neocortical connections, all else being equal.

2. We further suggest that generalized replay from a recurrent hippocampal system may also reduce the computational workload that the neocortex has to perform in order to extract the structure of the environment. Viewed from a geometric perspective, the process of consolidation may be thought of as involving the discovery of lower dimensional components (or manifolds) that underlie a set of related episodic experiences that occupy a high dimensional space. While we emphasize that generalization in REMERGE emerges at retrieval, we tentatively suggest that the

operation of recurrency (e.g., during offline periods) may also play a role in deriving more efficient lower dimensional representations of the underlying environmental structure. In the transitive inference task, for example, the goal would be to discover that the items presented during training form a linear structure (e.g., A–B–C–D–E–F). In this respect, it is worth noting similarities between the mechanism of generalization in REMERGE and the workings of the ISOMAP algorithm, a model that has been shown to increase the power of similarity based techniques (e.g., multidimensional scaling) in discovering the low dimensional structure of data sets (Tenenbaum, de Silva, & Langford, 2000). Both models provide a means by which the inherent similarity of two related experiences (e.g., B–C, D–E in the transitive inference task) can be captured, even when pairwise similarities in input-space are uninformative (and in this case equate to zero). REMERGE achieves this through a constraint satisfaction style search of the high dimensional space of episodes performed at retrieval, while the ISOMAP algorithm indexes the similarity of a pair of points by computing the shortest path between them. In the future, it will be interesting to explore whether REMERGE may offer a neurally inspired mechanism by which to learn compact representations of environmental structures.

The mechanism of hippocampal generalization we propose, however, also has important limitations of its own, suggesting it acts in concert with a slow learning neocortical system to support efficient semantic learning and semantic task performance.

1. The hippocampal system may support generalization over relatively short timescales. However, limitations on hippocampal storage capacity, due to its limited size, as well as a wealth of empirical evidence from patients with amnesia and semantic dementia, suggest that semantic knowledge is ultimately consolidated to the neocortex (McClelland et al., 1995; McClelland & Rogers, 2003; Rogers & McClelland, 2004; Scoville & Milner, 1957; Squire, 1992; Squire et al., 1984).

2. Our account emphasizes the principle that individual experiences are represented in the hippocampal system as arbitrary combinations of recurring components (or features), for example, denoting familiar objects, people, odors, or places (Cohen & Eichenbaum, 1993; McClelland & Goddard, 1996; McClelland et al., 1995). We view the neocortex as critical to the gradual development of neural representations that are shaped by the organizational structure of the environment and capture the general characteristics of our experiences (Rogers & McClelland, 2004). One very important characteristic of neocortical learning in this context is the potential for the gradual discovery of useful features that can then be used in rapid hippocampal learning. For example, consider learning a new word—for example, *bardrel* meaning (say) a kind of bird that lives in Patagonia. If this word is represented in the input to the hippocampus in terms of the specific line segments that make it up, and then a hippocampal representation is formed conjoining the elements of this visual pattern, the basis for generalization to other items will be limited to transformations that preserve the visual feature pattern. However, if this word is represented in the input to the hippocampus as a particular configuration of abstract letters, it will be possible to recognize the same word in distinct versions of the alphabet, including uppercase BARDREL or handwritten script. The development of a system of connection weights that assigns distinct visual patterns (such as a, *a*, and A) the same internal representation (corresponding to the letter A regardless of the particular graphical depiction) is a crucial function for an intelligent learning system and is thought to arise through gradual neocortical learning (e.g., McClelland & Goddard, 1996). Of note, such learning can assign representations to items corresponding to positions in a multi-dimensional space based only on their relationship to other items. These relationships may then be very abstract conceptual relationships (e.g., father-of, husband-of; Hinton et al., 1986). Just such a set of relationships is likely to be the basis on which the neocortex would learn the abstract correspondences among graphical variants of different letters.

3. It is important to bear in mind that generalization arises through very different mechanisms in a recurrent hippocampal system, compared to network models of the neocortical system. Specifically, the similarity structure of the environment is not statically represented in the internal representations of the network we describe, as it is in networks used to simulate neocortical learning (Rogers & McClelland, 2004), but is instead created on the fly through recurrence. While this property can be useful in affording considerable flexibility in response to changing environments (see below), the representation of regularities present in relatively stationary environments (e.g., familiar concepts, letter forms, objects, etc.) would appear to be most efficiently captured by the stable neural codes viewed to operate in the neocortex. Further, the variability of network output as a function of the level of inhibition, regulated by network temperature, a free parameter in the model, raises the question of how this parameter is controlled, an important issue that deserves consideration in future work.

One scenario that pinpoints a limitation of hippocampal generalization is where successful performance depends on exploiting the shared relational structure between two domains that are perceptually unrelated, as is the case in certain analogical reasoning tasks. The proposed hippocampal scheme would be expected to fail in this context because it lacks any mechanism that would allow transfer from one domain to another, assuming the absence of any cross-domain similarity in the inputs. Recent work has demonstrated that slower-learning models of the neocortex perform well in these settings, for example appropriately generalizing information learnt about one family to another family with an isomorphic relational structure (but no overlap in the inputs and outputs), through learning hidden layer representations that capture the shared structure across domains in the data set (Flusberg, Thibodeau, Sternberg, & Glick, 2011; Hinton et al., 1986; McClelland & Rogers, 2003; Rogers & McClelland, 2008). Such representations could then be exploited for new learning via the hippocampal system that could potentially generalize across domains. More work exploring how hippocampal and neocortical learning systems may work together in such settings is needed.

4. It has recently been shown in an exciting series of experimental studies in rodents that new information that is highly consistent with established neocortical knowledge structures can be rapidly integrated into cortical networks, requiring hippocampal involvement for only a short period (Tse et al., 2007, 2011). These recent findings indicate yet another way in which structured knowledge represented in the neocortex can complement the role of the hippocampus, and they underscore the importance of consistency with existing knowledge structures for allowing rapid neocortical integration. (Although these articles suggest that these

new findings might require modification of the complementary learning systems theory, simulations to be reported elsewhere by McClelland, 2011, now demonstrate that in fact such findings are consistent with the characteristics of the existing implementation used to illustrate many of the features of the neocortical learning system, namely that used by Rumelhart & Todd, 1993, and Rogers & McClelland, 2004.)

Taken together, we suggest that the inherent advantages (fast learning, flexibility, facilitates interleaved training) and limitations (dependence on pre-established components and structural relations, variable generalization as a function of inhibition, limited storage capacity) of a recurrence-based hippocampal scheme are largely mirrored by inverse advantages (gradually learnable componential codes that can capture structured relational abstractions, reliable generalization, large storage capacity) and disadvantages (slow learning, requires interleaved training) of neocortical networks. Overall, new developments continue to enrich our understanding of the ways in which complementary action of these two neural systems supports learning and memory, including generalization of what has been learned from new information.

Comment on Constructive Memory: Imagination, Future Thinking

We conclude with a speculation on how the perspective we advocate may have wider implications concerning the contribution of the hippocampus to cognition. A resurgent theme, first popularized by Barlett in the 1930s (Bartlett, 1932), is that memory is constructive in nature, rather than a literal record of the past (Hassabis & Maguire, 2007; McClelland, 2011; Nystrom & McClelland, 1992; Schacter & Addis, 2007). Despite rendering a memory system more prone to errors, it is this constructive aspect of memory that may be its most adaptive, facilitating the creation of new imagined scenarios through the recombination of multiple past experiences, thereby affording a capacity for the simulation of plausible future events. Interestingly, the hippocampus would appear to be critical to this form of generalization, based on evidence that the imagined creations of patients with amnesia are impoverished compared to control subjects (Hassabis, Kumaran, Vann, & Maguire, 2007). The mechanism by which the hippocampus achieves this function—and whether this reflects a specific role in spatial processing (Hassabis & Maguire, 2007; O'Keefe & Nadel, 1978) or a broader role in relational memory binding (Cohen & Eichenbaum, 1993)—remains unclear.

Here, we suggest that the incorporation of recurrence in the computations performed by the hippocampal system naturally affords constructive properties. Interestingly, previous work has shown that the tendency of subjects to show blending errors at the stage of memory recall (i.e., recombination of different sentences in a completion task) can be captured through the operation of a stochastic interactive process similar to the recurrent mechanism implemented in the REMERGE model (Nystrom & McClelland, 1992). Further, introspection on our own imagined creations, and even our dreams, would seem to support a simple intuition: Novel experiences typically involve familiar components (e.g., objects, people), or environmental regularities, rather than entirely fictitious items. Our suggestion, then, is that hippocampal recurrency, in combination with the compositional structure of memories implemented in our model and outlined in other viewpoints (Cohen & Eichenbaum, 1993; McClelland & Goddard, 1996; also see Hummel & Holyoak, 2003), is critical to supporting the arbitrary recombination of elements taken from multiple related episodes that underpins imagination, dreaming, and constructive memory more generally. Indeed, a recurrent memory system, including conjunctive representations that link co-occurring elements composing specific events, combined with top-down inputs from the prefrontal cortex, and the potential for the dynamic variation of network output determined by the level of inhibitory influences, would appear to offer considerable power and flexibility of the kind that may be important for imagination and future thinking.

Future Directions

We have aimed to capture key functional properties of the hippocampal system as a whole, following a simplified scheme designed to reveal how generalization may emerge through a principle of recurrent similarity computation. In future work it will be important to enhance the biological and functional realism of the model by including dynamic learning algorithms and intrinsic variability, and to capture the distinct contributions of anatomically defined brain structures within the hippocampal system, building on earlier extensions of the complementary learning systems framework (McClelland & Goddard, 1996; Norman & O'Reilly, 2003; O'Reilly & Rudy, 2001). These extensions to the model would also allow one to assess whether the core principle we propose, namely recurrent similarity computation, has broader applicability beyond the relatively simple network architectures and experimental scenarios explored here, to generalization in more complex environments.

To achieve this, it will be necessary to demonstrate that desirable functional properties of the REMERGE model (e.g., generalization capacity) can be retained in the context of learned sparse distributed representations of the nature presumed to exist within the hippocampus, an important challenge that remains to be fully addressed by the wider field. Here, we outline some of the key challenges for this endeavor.

In the current formulation, we examined the effects of a highly idealized version of learning on the network's capacity for generalization. Learning was simulated by stipulating the creation of new units for each premise item, then simply varying the relevant connection weights across different runs of the network to capture different levels of learning. Future work will be needed to employ dynamic simulations of a more realistic learning process to explore several issues, including the ability of the network to learn representations of the premise items themselves and the influence of the operation of recurrency during learning on representations in the network. This work should also assess the tendency of the network to learn representations of stored generalizations (e.g., linking different faces paired with the same scene in different episodes) that arise on the feature layer arising through recurrency—and the effect of these processes on the capacity of the network to generalize and yet retain the ability to recall specific experiences.

Here, we have used localist representations in the conjunctive layer, designed to reflect a theoretical extreme idealization of pattern separation in the hippocampus. In fact, however, our theory holds that the hippocampus supports neural representations that are sparser than in other brain regions (e.g., the entorhinal cortex), but that are nevertheless still distributed in nature. In the future, it will

be crucial to the further development of the theory to simulate this property in the model, with neural codes in the hidden layer rendered sparser than those in the feature layer through principles outlined in prior formulations—that is, competitive dynamics based on conjunctive coding, inhibitory influences, and low activity levels (e.g., Norman & O'Reilly, 2003; O'Reilly & McClelland, 1994; O'Reilly & Rudy, 2001).

As part of the current work, we have provided initial evidence that recurrency in the hippocampal circuit does not have inevitable costs for recognition memory. In future work, it will be important to examine this issue in more detail—in the context of a recurrent hippocampal system, instantiated using distributed neural codes and incorporating intrinsic variability in processing (i.e., noise). In particular, it would be illuminating to examine the characteristics of the recall signal triggered when study and lure items are presented to the recurrent network—and to compare these distributions to the performance of a feedforward-only hippocampal model (Norman and O'Reilly, 2003) and empirical data concerning the contribution of the hippocampus to recognition memory (e.g., Fortin, Wright, & Eichenbaum, 2004; Wais, Wixted, Hopkins, Squire, 2006). A further important development for this work will be to understand how biologically realistic mechanisms of learning can capture more fully the characteristics of cognitive-level differentiation models of recognition memory (Criss & McClelland, 2006; McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997).

We also hope to explore the possibility that a hippocampal system endowed with recurrent capabilities may facilitate the discovery of environmental structure and its representation in the neocortex, through the replay of generalized activity patterns during offline periods such as sleep. Further research is clearly needed to more fully explore this hypothesis: In broad terms, it would be interesting to assess whether generalized replay arising from a recurrent hippocampal system enhances the efficiency of neocortical learning, compared to the replay of individual episodes as envisioned in the original formulation of complementary learning systems theory.

More generally, we wish to highlight the need to develop a fuller and more systematic treatment of the range of different mechanisms that contribute to what we see as the cardinal aspect of knowledge acquisition, the ability to extract and exploit the general structure of a set of related experiences. Recurrent similarity-based generalization as envisioned here, together with more gradual structure learning as originally envisioned in the complementary learning systems theory, may also co-exist with other forms of semantic learning, including the use of explicit reasoning strategies and the formation of explicitly structured representations such as linear orderings. This understanding should also explore ways in which the processes that lead to these generalizations would still allow retrieval of the particulars of individual events and experiences. The research we present here will hopefully contribute, in part, to the development of such a systematic understanding.

Conclusion

In this article, we have drawn attention to a fundamental, but neglected, tension between theories emphasizing the role of the hippocampus in pattern separation and episodic memory (Marr, 1971; McClelland et al., 1995) and those highlighting its contribution to generalization, often known as flexible or relational memory (Cohen & Eichenbaum, 1993). We have proposed a means by which this apparent conflict may be resolved, through the exploitation of recurrent similarity-based computation in the hippocampal system. This proposal allows a theoretical ideal of pattern separated representations to be retained, while also allowing for efficient generalization. The explanatory value of the REMERGE model will surely rest on future empirical and theoretical developments—at a minimum, we hope our perspective will stimulate discussion and will provoke further investigation of issues that speak to the essential characteristics of learning and memory in the brain.

References

Amaral, D. G., & Lavenex, P. (2006). Hippocampal neuroanatomy. In T. Bliss, P. Andersen, D. G. Amaral, R. G. Morris, & J. O'Keefe (Eds.), *The hippocampus book* (pp. 37–115). Oxford, England: Oxford University Press.

Ashby, F. G., & Maddox, W. T. (2005). Human category learning. *Annual Review of Psychology, 56,* 149–178. doi:10.1146/annurev.psych.56.091103.070217

Bartlett, F. C. (1932). *Remembering: An experimental and social study.* Cambridge, England: Cambridge University Press.

Becker, S. (2005). A computational principle for hippocampal learning and neurogenesis. *Hippocampus, 15,* 722–738. doi:10.1002/hipo.20095

Bowers, J. S. (2009). On the biological plausibility of grandmother cells: Implications for neural network theories in psychology and neuroscience. *Psychological Review, 116,* 220–251. doi:10.1037/a0014462

Brawn, T. P., Nusbaum, H. C., & Margoliash, D. (2010). Sleep-dependent consolidation of auditory discrimination learning in adult starlings. *The Journal of Neuroscience, 30,* 609–613. doi:10.1523/JNEUROSCI.4237-09.2010

Breslow, L. (1981). A reevaluation of the literature on the development of transitive inferences. *Psychological Bulletin, 89,* 325–351. doi:10.1037/0033-2909.89.2.325

Brown, M. W., & Aggleton, J. P. (2001). Recognition memory: What are the roles of the perirhinal cortex and hippocampus? *Nature Reviews Neuroscience, 2,* 51–61. doi:10.1038/35049064

Buckmaster, C. A., Eichenbaum, H., Amaral, D. G., Suzuki, W. A., & Rapp, P. R. (2004). Entorhinal cortex lesions disrupt the relational organization of memory in monkeys. *The Journal of Neuroscience, 24,* 9811–9825. doi:10.1523/JNEUROSCI.1532-04.2004

Bunsey, M., & Eichenbaum, H. (1996, January 18). Conservation of hippocampal memory function in rats and humans. *Nature, 379,* 255–257. doi:10.1038/379255a0

Burgess, N. (2006). Computational models of the spatial and mnemonic functions of the hippocampus. In T. Bliss, P. Andersen, D. G. Amaral, R. G. Morris, & J. O'Keefe (Eds.), *The hippocampus book* (pp. 715–751). Oxford, England: Oxford University Press.

Buzsáki, G. (1986). Hippocampal sharp waves: Their origin and significance. *Brain Research, 398,* 242–252. doi:10.1016/0006-8993(86)91483-6

Buzsáki, G. (1989). Two-stage model of memory trace formation: A role for "noisy" brain states. *Neuroscience, 31,* 551–570. doi:10.1016/0306-4522(89)90423-5

Buzsáki, G. (1996). The hippocampo-neocortical dialogue. *Cerebral Cortex, 6,* 81–92. doi:10.1093/cercor/6.2.81

Buzsáki, G. (2005). Theta rhythm of navigation: Link between path integration and landmark navigation, episodic and semantic memory. *Hippocampus, 15,* 827–840. doi:10.1002/hipo.20113

Cai, D. J., Mednick, S. A., Harrison, E. M., Kanady, J. C., & Mednick, S. C. (2009). REM, not incubation, improves creativity by priming associative networks. *Proceedings of the National Academy of Sciences, USA, 106,* 10130–10134. doi:10.1073/pnas.0900271106

Clelland, C. D., Choi, M., Romberg, C., Clemenson, G. D., Jr., Fragniere,

A., Tyers, P., . . . Bussey, T. J. (2009, July 10). A functional role for adult hippocampal neurogenesis in spatial pattern separation. *Science, 325,* 210–213. doi:10.1126/science.1173215

Cohen, N. J., & Eichenbaum, H. (1993). *Memory, amnesia, and the hippocampal system.* Cambridge, MA: MIT Press.

Coutureau, E., Killcross, A. S., Good, M., Marshall, V. J., Ward-Robinson, J., & Honey, R. C. (2002). Acquired equivalence and distinctiveness of cues: II. Neural manipulations and their implications. *Journal of Experimental Psychology: Animal Behavior Processes, 28,* 388–396. doi:10.1037/0097-7403.28.4.388

Criss, A. H., & McClelland, J. L. (2006). Differentiating the differentiation models: A comparison of the retrieving effectively from memory model (REM) and the subjective likelihood model (SLiM). *Journal of Memory and Language, 55,* 447–460. doi:10.1016/j.jml.2006.06.003

Davis, H. (1992). Transitive inference in rats (*Rattus norvegicus*). *Journal of Comparative Psychology, 106,* 342–349. doi:10.1037/0735-7036.106.4.342

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience, 8,* 1704–1711. doi:10.1038/nn1560

Delius, J. D., & Sieman, M. (1998). Transitive responding in animals and humans: Exaptation rather than adaptation. *Behavioural Processes, 42,* 107–137. doi:10.1016/S0376-6357(97)00072-7

Deng, W., Aimone, J. B., & Gage, F. H. (2010). New neurons and new memories: How does adult hippocampal neurogenesis affect learning and memory? *Nature Reviews Neuroscience, 11,* 339–350. doi:10.1038/nrn2822

Desimone, R. (1998). Visual attention mediated by biased competition in extrastriate visual cortex. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 353,* 1245–1255. doi:10.1098/rstb.1998.0280

Diekelmann, S., & Born, J. (2010). The memory function of sleep. *Nature Reviews Neuroscience, 11,* 114–126.

Duncan, J. (2001). An adaptive coding model of neural function in prefrontal cortex. *Nature Reviews Neuroscience, 2,* 820–829. doi:10.1038/35097575

Dusek, J. A., & Eichenbaum, H. (1997). The hippocampus and memory for orderly stimulus relations. *Proceedings of the National Academy of Sciences, USA, 94,* 7109–7114. doi:10.1073/pnas.94.13.7109

Ego-Stengel, V., & Wilson, M. A. (2010). Disruption of ripple-associated hippocampal activity during rest impairs spatial learning in the rat. *Hippocampus, 20,* 1–10.

Eichenbaum, H. (1999). The hippocampus and mechanisms of declarative memory. *Behavioural Brain Research, 103,* 123–133. doi:10.1016/S0166-4328(99)00044-3

Eichenbaum, H. (2004). Hippocampus: Cognitive processes and neural representations that underlie declarative memory. *Neuron, 44,* 109–120. doi:10.1016/j.neuron.2004.08.028

Eichenbaum, H., Dudchenko, P., Wood, E., Shapiro, M., & Tanila, H. (1999). The hippocampus, memory, and place cells: Is it spatial memory or a memory space? *Neuron, 23,* 209–226. doi:10.1016/S0896-6273(00)80773-4

Eichenbaum, H., Stewart, C., & Morris, R. G. (1990). Hippocampal representation in place learning. *Journal of Neuroscience, 10,* 3531–3542.

Eichenbaum, H., Yonelinas, A. P., & Ranganath, C. (2007). The medial temporal lobe and recognition memory. *Annual Review of Neuroscience, 30,* 123–152. doi:10.1146/annurev.neuro.30.051606.094328

Ellenbogen, J. M., Hu, P. T., Payne, J. D., Titone, D., & Walker, M. P. (2007). Human relational memory requires time and sleep. *Proceedings of the National Academy of Sciences, USA, 104,* 7723–7728. doi:10.1073/pnas.0700094104

Ferbinteanu, J., & Shapiro, M. L. (2003). Prospective and retrospective memory coding in the hippocampus. *Neuron, 40,* 1227–1239. doi:10.1016/S0896-6273(03)00752-9

Flusberg, S. J., Thibodeau, P. H., Sternberg, D. A., & Glick, J. J. (2011). A connectionist approach to embodied conceptual metaphor. *Frontiers in Psychology, 1,* 197.

Fortin, N. J., Wright, S. P., & Eichenbaum, H. (2004, September 9). Recollection-like memory retrieval in rats is dependent on the hippocampus. *Nature, 431,* 188–191. doi:10.1038/nature02853

Foster, D. J., & Wilson, M. A. (2006, March 30). Reverse replay of behavioural sequences in hippocampal place cells during the awake state. *Nature, 440,* 680–683. doi:10.1038/nature04587

Frank, M. J., Rudy, J. W., Levy, W. B., & O'Reilly, R. C. (2005). When logic fails: Implicit transitive inference in humans. *Memory & Cognition, 33,* 742–750. doi:10.3758/BF03195340

Frank, M. J., Rudy, J. W., & O'Reilly, R. C. (2003). Transitivity, flexibility, conjunctive representations, and the hippocampus: II. A computational analysis. *Hippocampus, 13,* 341–354. doi:10.1002/hipo.10084

French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences, 3,* 128–135. doi:10.1016/S1364-6613(99)01294-2

Gilbert, P. E., Kesner, R. P., & Lee, I. (2001). Dissociating hippocampal subregions: Double dissociation between dentate gyrus and CA1. *Hippocampus, 11,* 626–636. doi:10.1002/hipo.1077

Gillan, D. J. (1981). Reasoning in the chimpanzee: II. Transitive inference. *Journal of Experimental Psychology: Animal Behavior Processes, 7,* 150–164. doi:10.1037/0097-7403.7.2.150

Girardeau, G., Benchenane, K., Wiener, S. I., Buzsáki, G., & Zugaro, M. B. (2009). Selective suppression of hippocampal ripples impairs spatial memory. *Nature Neuroscience, 12,* 1222–1223. doi:10.1038/nn.2384

Gluck, M. A., & Myers, C. E. (1993). Hippocampal mediation of stimulus representation: A computational theory. *Hippocampus, 3,* 491–516. doi:10.1002/hipo.450030410

Greene, A. J., Spellman, B. A., Dusek, J. A., Eichenbaum, H. B., & Levy, W. B. (2001). Relational learning with and without awareness: Transitive inference using nonverbal stimuli in humans. *Memory & Cognition, 29,* 893–902. doi:10.3758/BF03196418

Grossberg, S. (1978). A theory of visual coding, memory, and development. In E. L. J. Leeuwenberg & H. F. J. M. Buffart (Eds.), *Formal theories of visual perception* (pp. 7–26). New York, NY: Wiley.

Grossberg, S. (1987). Competitive learning: From interactive activation to adaptive resonance. *Cognitive Science, 11,* 23–63. doi:10.1111/j.1551-6708.1987.tb00862.x

Gupta, A. S., van der Meer, M. A., Touretzky, D. S., & Redish, A. D. (2010). Hippocampal replay is not a simple function of experience. *Neuron, 65,* 695–705. doi:10.1016/j.neuron.2010.01.034

Hassabis, D., Kumaran, D., Vann, S. D., & Maguire, E. A. (2007). Patients with hippocampal amnesia cannot imagine new experiences. *Proceedings of the National Academy of Sciences, USA, 104,* 1726–1731. doi:10.1073/pnas.0610561104

Hassabis, D., & Maguire, E. A. (2007). Deconstructing episodic memory with construction. *Trends in Cognitive Sciences, 11,* 299–306. doi:10.1016/j.tics.2007.05.001

Hasselmo, M. E. (1993). Acetylcholine and learning in a cortical associative memory network. *Neural Computation, 5,* 32–44. doi:10.1162/neco.1993.5.1.32

Hasselmo, M. E. (1999). Neuromodulation: Acetylcholine and memory consolidation. *Trends in Cognitive Sciences, 3,* 351–359. doi:10.1016/S1364-6613(99)01365-0

Hasselmo, M. E. (2005). What is the function of hippocampal theta rhythm? Linking behavioral data to phasic properties of field potential and unit recording data. *Hippocampus, 15,* 936–949. doi:10.1002/hipo.20116

Hasselmo, M. E., & Bower, J. M. (1992). Cholinergic suppression specific to intrinsic not afferent fiber synapses in rat piriform (olfactory) cortex. *Journal of Neurophysiology, 67,* 1222–1229.

Hasselmo, M. E., & Schnell, E. (1994). Laminar selectivity of the cholinergic suppression of synaptic transmission in rat hippocampal region

CA1: Computational modeling and brain slice physiology. *Journal of Neuroscience, 14,* 3898–3914.

Heckers, S., Zalesak, M., Weiss, A. P., Ditman, T., & Titone, D. (2004). Hippocampal activation during transitive inference in humans. *Hippocampus, 14,* 153–162. doi:10.1002/hipo.10189

Hinton, G. E., McClelland, J. L., & Rumelhart, D. E. (1986). Distributed representations. In J. L. McClelland, D. E. Rumelhart, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition: Psychological and biological models* (Vol. 2, pp. 77–109). Cambridge, MA: MIT Press.

Hinton, G. E., & Sejnowksi, T. J. (1983). Optimal perceptual inference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 448–453). Retrieved from http://www.stanford.edu/~ngoodman/Hinton_Sejnowski_OptimalPerceptualDifference_1983.pdf

Hintzman, D. L. (1986). "Schema abstraction" in a multiple-trace memory model. *Psychological Review, 93,* 411–428. doi:10.1037/0033-295X.93.4.411

Hintzman, D. L. (2001). Similarity, global matching, and judgments of frequency. *Memory & Cognition, 29,* 547–556. doi:10.3758/BF03200456

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences, USA, 79,* 2554–2558. doi:10.1073/pnas.79.8.2554

Howard, M. W., Fotedar, M. S., Datey, A. V., & Hasselmo, M. E. (2005). The temporal context model in spatial navigation and relational learning: Toward a common explanation of medial temporal lobe function across domains. *Psychological Review, 112,* 75–116. doi:10.1037/0033-295X.112.1.75

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review, 110,* 220–264. doi:10.1037/0033-295X.110.2.220

Hunsaker, M. R., Rosenberg, J. S., & Kesner, R. P. (2008). The role of the dentate gyrus, CA3a, b, and CA3c for detecting spatial and environmental novelty. *Hippocampus, 18,* 1064–1073. doi:10.1002/hipo.20464

Ji, D., & Wilson, M. A. (2007). Coordinated memory replay in the visual cortex and hippocampus during sleep. *Nature Neuroscience, 10,* 100–107. doi:10.1038/nn1825

Johnson, A., & Redish, A. D. (2005). Hippocampal replay contributes to within session learning in a temporal difference reinforcement learning model. *Neural Networks, 18,* 1163–1171. doi:10.1016/j.neunet.2005.08.009

Kahana, M. J. (1996). Associative retrieval processes in free recall. *Memory & Cognition, 24,* 103–109. doi:10.3758/BF03197276

Káli, S., & Dayan, P. (2004). Off-line replay maintains declarative memories in a model of hippocampal-neocortical interactions. *Nature Neuroscience, 7,* 286–294. doi:10.1038/nn1202

Kempermann, G. (2002). Why new neurons? Possible functions for adult hippocampal neurogenesis. *The Journal of Neuroscience, 22,* 635–638.

Kesner, R. P., & Hopkins, R. O. (2006). Mnemonic functions of the hippocampus: A comparison between animals and humans. *Biological Psychology, 73,* 3–18. doi:10.1016/j.biopsycho.2006.01.004

Kitchener, E. G., & Squire, L. R. (2000). Impaired verbal category learning in amnesia. *Behavioral Neuroscience, 114,* 907–911. doi:10.1037/0735-7044.114.5.907

Kloosterman, F., van Haeften, T., & Lopes da Silva, F. H. (2004). Two reentrant pathways in the hippocampal-entorhinal system. *Hippocampus, 14,* 1026–1039. doi:10.1002/hipo.20022

Knowlton, B. J., & Squire, L. R. (1993, December 10). The learning of categories: Parallel brain systems for item memory and category knowledge. *Science, 262,* 1747–1749. doi:10.1126/science.8259522

Kumaran, D., Hassabis, D., Spiers, H. J., Vann, S. D., Vargha-Khadem, F., & Maguire, E. A. (2007). Impaired spatial and non-spatial configural learning in patients with hippocampal pathology. *Neuropsychologia, 45,* 2699–2711. doi:10.1016/j.neuropsychologia.2007.04.007

Kumaran, D., Summerfield, J. J., Hassabis, D., & Maguire, E. A. (2009). Tracking the emergence of conceptual knowledge during human decision making. *Neuron, 63,* 889–901. doi:10.1016/j.neuron.2009.07.030

Lee, A. K., & Wilson, M. A. (2002). Memory of sequential experience in the hippocampus during slow wave sleep. *Neuron, 36,* 1183–1194. doi:10.1016/S0896-6273(02)01096-6

Leutgeb, J. K., Leutgeb, S., Moser, M. B., & Moser, E. I. (2007, February 16). Pattern separation in the dentate gyrus and CA3 of the hippocampus. *Science, 315,* 961–966. doi:10.1126/science.1135801

Leutgeb, S., Leutgeb, J. K., Treves, A., Moser, M. B., & Moser, E. I. (2004, August 27). Distinct ensemble codes in hippocampal areas CA3 and CA1. *Science, 305,* 1295–1298. doi:10.1126/science.1100265

Lisman, J. E. (1999). Relating hippocampal circuitry to function: Recall of memory sequences by reciprocal dentate-CA3 interactions. *Neuron, 22,* 233–242. doi:10.1016/S0896-6273(00)81085-5

Luce, R. D. (1959). *Individual choice behavior: A theoretical analysis.* New York, NY: Wiley.

Marr, D. (1971). Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 262,* 23–81. doi:10.1098/rstb.1971.0078

McClelland, J. L. (1981). Retrieving general and specific information from stored knowledge of the specifics. In *Proceedings of the Third Annual Conference of the Cognitive Science Society* (pp. 170–172). Retrieved from http://psych.stanford.edu/~jlm/papers/McClelland81.pdf

McClelland, J. L. (1991). Stochastic interactive processes and the effect of context on perception. *Cognitive Psychology, 23,* 1–44. doi:10.1016/0010-0285(91)90002-6

McClelland, J. L. (1998). Connectionist models and Bayesian inference. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (pp. 21–53). Oxford, England: Oxford University Press.

McClelland, J. L. (2011). Memory as a constructive process: The parallel-distributed processing approach. In P. Nalbantian, P. Matthew, & J. L. McClelland (Eds.), *The memory process: Neuroscientific and humanist perspectives* (pp. 129–152). Cambridge, MA: MIT Press.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in Cognitive Sciences, 14,* 348–356. doi:10.1016/j.tics.2010.06.002

McClelland, J. L., & Chappell, M. (1998). Familiarity breeds differentiation: A subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review, 105,* 724–760. doi:10.1037/0033-295X.105.4.734-760

McClelland, J. L., & Goddard, N. H. (1996). Considerations arising from a complementary learning systems perspective on hippocampus and neocortex. *Hippocampus, 6,* 654–665. doi:10.1002/(SICI)1098-1063(1996)6:6<654::AID-HIPO8>3.0.CO;2-G

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102,* 419–457. doi:10.1037/0033-295X.102.3.419

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience, 4,* 310–322. doi:10.1038/nrn1076

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of the basic findings. *Psychological Review, 88,* 375–407. doi:10.1037/0033-295X.88.5.375

McCloskey, M., & Cohen, N. J. (1989). Catastrophic forgetting in connectionist networks: The problem of sequential learning. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 20, pp. 109–165). New York, NY: Academic Press.

McGonigle, B. O., & Chalmers, M. (1977, June 23). Are monkeys logical? *Nature, 267,* 694–696. doi:10.1038/267694a0

McHugh, T. J., Jones, M. W., Quinn, J. J., Balthasar, N., Coppari, R., Elmquist, J. K., . . . Tonegawa, S. (2007, July 6). Dentate gyrus NMDA receptors mediate rapid pattern separation in the hippocampal network. *Science, 317,* 94–99. doi:10.1126/science.1140263

McNaughton, B. L., & Morris, R. G. (1987). Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends in Neurosciences, 10,* 408–415. doi:10.1016/0166-2236(87)90011-7

Medin, D. L. (1975). A theory of context in discrimination learning. In G. H. Bower (Ed.), *Psychology of learning and motivation* (Vol. 9, pp. 269–315). New York, NY: Academic Press.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification. *Psychological Review, 85,* 207–238. doi:10.1037/0033-295X.85.3.207

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual Review of Neuroscience, 24,* 167–202. doi:10.1146/annurev.neuro.24.1.167

Mirman, D., Khaitan, P., Bolger, D. J., & McClelland, J. L. (in press). Interactive activation and mutual constraint satisfaction in perception and cognition. *Cognitive Science.*

Moita, M. A., Rosis, S., Zhou, Y., LeDoux, J. E., & Blair, H. T. (2003). Hippocampal place cells acquire location-specific responses to the conditioned stimulus during auditory fear conditioning. *Neuron, 37,* 485–497. doi:10.1016/S0896-6273(03)00033-3

Molter, C., Sato, N., & Yamaguchi, Y. (2007). Reactivation of behavioral activity during sharp waves: A computational model for two stage hippocampal dynamics. *Hippocampus, 17,* 201–209. doi:10.1002/hipo.20258

Moses, S. N., Villate, C., & Ryan, J. D. (2006). An investigation of learning strategy supporting transitive inference performance in humans compared to other species. *Neuropsychologia, 44,* 1370–1387. doi:10.1016/j.neuropsychologia.2006.01.004

Murphy, G. L. (2004). *The big book of concepts.* Cambridge, MA: MIT Press.

Murphy, R. A., Mondragon, E., & Murphy, V. A. (2008, March 28). Rule learning by rats. *Science, 319,* 1849–1851. doi:10.1126/science.1151564

Myers, C. E., & Scharfman, H. E. (2011). Pattern separation in the dentate gyrus: A role for the CA3 backprojection. *Hippocampus, 21,* 1190–1215. doi:10.1002/hipo.20828

Myers, C. E., Shohamy, D., Gluck, M. A., Grossman, S., Kluger, A., Ferris, S., . . . Schwartz, R. (2003). Dissociating hippocampal versus basal ganglia contributions to learning and transfer. *Journal of Cognitive Neuroscience, 15,* 185–193. doi:10.1162/089892903321208123

Nakashiba, T., Cushman, J. D., Pelkey, K. A., Renaudineau, S., Buhl, D. L., McHugh, T. J., . . . Tonegawa, S. (2012). Young dentate granule cells mediate pattern separation, whereas old granule cells facilitate pattern completion. *Cell, 149,* 188–201. doi:10.1016/j.cell.2012.01.046

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review, 110,* 611–646. doi:10.1037/0033-295X.110.4.611

Nosofsky, R. M. (1984). Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10,* 104–114. doi:10.1037/0278-7393.10.1.104

Nosofsky, R. M. (2000). Exemplar representation without generalization? Comment on Smith and Minda's "Thirty categorization results in search of a model." *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 1735–1743.

Nosofsky, R. M., Little, D. R., & James, T. W. (2012). Activation in the neural network responsible for categorization and recognition reflects parameter changes. *Proceedings of the National Academy of Sciences, USA, 109,* 333–338. doi:10.1073/pnas.1111304109

Nosofsky, R. M., & Zaki, S. R. (1998). Dissociations between categorization and recognition in amnesic and normal individuals: An exemplar-based interpretation. *Psychological Science, 9,* 247–255. doi:10.1111/1467-9280.00051

Nystrom, L. E., & McClelland, J. L. (1992). Trace synthesis in cued recall. *Journal of Memory and Language, 31,* 591–614. doi:10.1016/0749-596X(92)90030-2

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

O'Neill, J., Pleydell-Bouverie, B., Dupret, D., & Csicsvari, J. (2010). Play it again: Reactivation of waking experience and memory. *Trends in Neurosciences, 33,* 220–229. doi:10.1016/j.tins.2010.01.006

O'Reilly, R. C., & McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall: Avoiding a trade-off. *Hippocampus, 4,* 661–682. doi:10.1002/hipo.450040605

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

O'Reilly, R. C., & Rudy, J. W. (2001). Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psychological Review, 108,* 311–345. doi:10.1037/0033-295X.108.2.311

Page, M. (2000). Connectionist modelling in psychology: A localist manifesto. *Behavioral and Brain Sciences, 23,* 443–467. doi:10.1017/S0140525X00003356

Palmeri, T. J., & Flannery, M. A. (1999). Learning about categories in the absence of training: Profound amnesia and the relationship between perceptual categorization and recognition memory. *Psychological Science, 10,* 526–530. doi:10.1111/1467-9280.00200

Plaut, D. C., & McClelland, J. L. (2010). Locating object knowledge in the brain: Comment on Bowers's (2009) attempt to revive the grandmother cell hypothesis. *Psychological Review, 117,* 284–288. doi:10.1037/a0017101

Polyn, S. M., & Kahana, M. J. (2008). Memory search and the neural representation of context. *Trends in Cognitive Sciences, 12,* 24–30. doi:10.1016/j.tics.2007.10.010

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review, 116,* 129–156. doi:10.1037/a0014420

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology, 77,* 353–363. doi:10.1037/h0025953

Preston, A. R., Shrager, Y., Dudukovic, N. M., & Gabrieli, J. D. (2004). Hippocampal contribution to the novel use of relational information in declarative memory. *Hippocampus, 14,* 148–152. doi:10.1002/hipo.20009

Quiroga, R. Q., Kreiman, G., Koch, C., & Fried, I. (2008). Sparse but not "grandmother-cell" coding in the medial temporal lobe. *Trends in Cognitive Sciences, 12,* 87–91. doi:10.1016/j.tics.2007.12.003

Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C., & Fried, I. (2005, June 23). Invariant visual representation by single neurons in the human brain. *Nature, 435,* 1102–1107. doi:10.1038/nature03687

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach.* Cambridge, MA: MIT Press.

Rogers, T. T., & McClelland, J. L. (2008). Précis of semantic cognition: A parallel distributed processing approach. *Behavioral and Brain Sciences, 31,* 689–714. doi:10.1017/S0140525X0800589X

Rolls, E. T., & Kesner, R. P. (2006). A computational theory of hippocampal function, and empirical tests of the theory. *Progress in Neurobiology, 79,* 1–48. doi:10.1016/j.pneurobio.2006.04.005

Rudy, J. W., & Sutherland, R. J. (1989). The hippocampal formation is necessary for rats to learn and remember configural discriminations. *Behavioural Brain Research, 34*(1–2), 97–109. doi:10.1016/S0166-4328(89)80093-2

Rumelhart, D. E. (1990). Brain style computation: Learning and generalization. In S. F. Zornetzer, J. L. Davis, & C. Lau (Eds.), *An introduction*

to electronic and neural networks (pp. 405–420). San Diego, CA: Academic Press.

Rumelhart, D. E., & Todd, P. M. (1993). Learning and connectionist representations. In D. M. Meyer & S. Kornblum (Eds.), *Attention and performance XIV: Synergies in experimental psychology* (pp. 3–30). Cambridge, MA: MIT Press.

Ryan, J. D., Moses, S. N., & Villate, C. (2009). Impaired relational organization of propositions, but intact transitive inference, in aging: Implications for understanding underlying neural integrity. *Neuropsychologia, 47,* 338–353. doi:10.1016/j.neuropsychologia.2008.09.006

Schacter, D. L., & Addis, D. R. (2007). The cognitive neuroscience of constructive memory: Remembering the past and imagining the future. *Philosophical Transactions of the Royal Society of London B: Biological Sciences, 362,* 773–786. doi:10.1098/rstb.2007.2087

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery & Psychiatry, 20,* 11–21. doi:10.1136/jnnp.20.1.11

Shepard, R. N. (1987, September 11). Toward a universal law of generalization for psychological science. *Science, 237,* 1317–1323. doi:10.1126/science.3629243

Shi, L., Griffiths, T. L., Feldman, N. H., & Sanborn, A. N. (2010). Exemplar models as a mechanism for performing Bayesian inference. *Psychonomic Bulletin & Review, 17,* 443–464. doi:10.3758/PBR.17.4.443

Shiffrin, R. M., & Steyvers, M. (1997). Models of recognition memory: REM-retrieving effectively from memory. *Psychonomic Bulletin & Review, 4,* 145–166. doi:10.3758/BF03209391

Shohamy, D., & Wagner, A. D. (2008). Integrating memories in the human brain: Hippocampal-midbrain encoding of overlapping events. *Neuron, 60,* 378–389. doi:10.1016/j.neuron.2008.09.023

Skaggs, W. E., & McNaughton, B. L. (1996, March 29). Replay of neuronal firing sequences in rat hippocampus during sleep following spatial experience. *Science, 271,* 1870–1873. doi:10.1126/science.271.5257.1870

Skaggs, W. E., & McNaughton, B. L. (1998). Spatial firing properties of hippocampal CA1 populations in an environment containing two visually identical regions. *The Journal of Neuroscience, 18,* 8455–8466.

Smith, C., & Squire, L. R. (2005). Declarative memory, awareness, and transitive inference. *The Journal of Neuroscience, 25,* 10138–10146. doi:10.1523/JNEUROSCI.2731-05.2005

Smith, J. D., & Minda, J. P. (2000). Thirty categorization results in search of a model. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 3–27. doi:10.1037/0278-7393.26.1.3

Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review, 99,* 195–231. doi:10.1037/0033-295X.99.2.195

Squire, L. R., Cohen, N. J., & Nadel, L. (1984). The medial temporal region and memory consolidation: A new hypothesis. In H. Weingartner & E. Parker (Eds.), *Memory consolidation* (pp. 185–210). Hillsdale, NJ: Erlbaum.

Squire, L. R., & Knowlton, B. J. (1995). Learning about categories in the absence of memory. *Proceedings of the National Academy of Sciences, USA, 92,* 12470–12474. doi:10.1073/pnas.92.26.12470

Squire, L. R., Stark, C. E., & Clark, R. E. (2004). The medial temporal lobe. *Annual Review of Neuroscience, 27,* 279–306. doi:10.1146/annurev.neuro.27.070203.144130

Stickgold, R., Scott, L., Rittenhouse, C., & Hobson, A. J. (1999). Sleep-induced changes in associative memory. *Journal of Cognitive Neuroscience, 11,* 182–193.

Tanila, H. (1999). Hippocampal place cells can develop distinct representations of two visually identical environments. *Hippocampus, 9,* 235–246. doi:10.1002/(SICI)1098-1063(1999)9:3<235::AID-HIPO4>3.0.CO;2-3

Tenenbaum, J. B., de Silva, V., & Langford, J. C. (2000, December 22). A global geometric framework for nonlinear dimensionality reduction. *Science, 290,* 2319–2323. doi:10.1126/science.290.5500.2319

Treves, A., & Rolls, E. T. (1992). Computational constraints suggest the need for two distinct input systems to the hippocampal CA3 network. *Hippocampus, 2,* 189–199. doi:10.1002/hipo.450020209

Treves, A., & Rolls, E. T. (1994). Computational analysis of the role of the hippocampus in memory. *Hippocampus, 4,* 374–391. doi:10.1002/hipo.450040319

Tse, D., Langston, R. F., Kakeyama, M., Bethus, I., Spooner, P. A., Wood, E. R., . . . Morris, R. G. M. (2007, April 6). Schemas and memory consolidation. *Science, 316,* 76–82. doi:10.1126/science.1135935

Tse, D., Takeuchi, T., Kakeyama, M., Kajii, Y., Okuno, H., Tohyama, C., . . . Morris, R. G. M. (2011, August 12). Schema-dependent gene activation and memory encoding in neocortex. *Science, 333,* 891–895. doi:10.1126/science.1205274

Tucker, M. A., & Fishbein, W. (2008). Enhancement of declarative memory performance following a daytime nap is contingent on strength of initial task acquisition. *Sleep, 31,* 197–203.

Usher, M., & McClelland, J. L. (2001). The time course of perceptual choice: The leaky, competing accumulator model. *Psychological Review, 108,* 550–592. doi:10.1037/0033-295X.108.3.550

Van Elzakker, M., O'Reilly, R. C., & Rudy, J. W. (2003). Transitivity, flexibility, conjunctive representations, and the hippocampus: I. An empirical analysis. *Hippocampus, 13,* 334–340. doi:10.1002/hipo.10083

van Strien, N. M., Cappaert, N. L., & Witter, M. P. (2009). The anatomy of memory: An interactive overview of the parahippocampal-hippocampal network. *Nature Reviews Neuroscience, 10,* 272–282. doi:10.1038/nrn2614

Vazdarjanova, A., & Guzowski, J. F. (2004). Differences in hippocampal neuronal population responses to modifications of an environmental context: Evidence for distinct, yet complementary, functions of CA3 and CA1 ensembles. *The Journal of Neuroscience, 24,* 6489–6496. doi:10.1523/JNEUROSCI.0350-04.2004

von Fersen, L., Wynne, C. D., Delius, J. D., & Staddon, J. E. (1991). Transitive inference formation in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes, 17,* 334–341. doi:10.1037/0097-7403.17.3.334

Wagner, A. D., Schacter, D. L., Rotte, M., Koutstaal, W., Maril, A., Dale, A. M., . . . Buckner, R. L. (1998, August 21). Building memories: Remembering and forgetting of verbal experiences as predicted by brain activity. *Science, 281,* 1188–1191. doi:10.1126/science.281.5380.1188

Wais, P. E., Wixted, J. T., Hopkins, R. O., & Squire, L. R. (2006). The Hippocampus supports both the recollection and the familiarity components of recognition memory. *Neuron, 49,* 459–466. doi:10.1016/j.neuron.2005.12.020

Walker, M. P., & Stickgold, R. (2010). Overnight alchemy: Sleep-dependent memory evolution. *Nature Reviews Neuroscience, 11,* 218. doi:10.1038/nrn2762–c1

Wamsley, E. J., Tucker, M., Payne, J. D., Benavides, J. A., & Stickgold, R. (2010). Dreaming of a learning task is associated with enhanced sleep-dependent memory consolidation. *Current Biology, 20,* 850–855. doi:10.1016/j.cub.2010.03.027

Waydo, S., Kraskov, A., Quian Quiroga, R., Fried, I., & Koch, C. (2006). Sparse representation in the human medial temporal lobe. *The Journal of Neuroscience, 26,* 10232–10234. doi:10.1523/JNEUROSCI.2101-06.2006

Wilson, M. A., & McNaughton, B. L. (1994, July 29). Reactivation of hippocampal ensemble memories during sleep. *Science, 265,* 676–679. doi:10.1126/science.8036517

Wixted, J. T. (2004). The psychology and neuroscience of forgetting. *Annual Review of Psychology, 55,* 235–269. doi:10.1146/annurev.psych.55.090902.141555

Wixted, J. T., & Squire, L. R. (2010). The role of the human hippocampus in familiarity-based and recollection-based recognition memory. *Behavioural Brain Research, 215,* 197–208. doi:10.1016/j.bbr.2010.04.020

Wood, E. R., Dudchenko, P. A., & Eichenbaum, H. (1999, February 18).

The global record of memory in hippocampal neuronal activity. *Nature, 397,* 613–616. doi:10.1038/17605

Wu, X., & Levy, W. B. (2001). Simulating symbolic distance effects in the transitive inference problem. *Neurocomputing, 38–40,* 1603–1610. doi:10.1016/S0925-2312(01)00512-4

Zaki, S. R. (2004). Is categorization performance really intact in amnesia? A meta-analysis. *Psychonomic Bulletin & Review, 11,* 1048–1054. doi:10.3758/BF03196735

Zeithamova, D., & Preston, A. R. (2010). Flexible memories: Differential roles for medial temporal lobe and prefrontal cortex in cross-episode binding. *The Journal of Neuroscience, 30,* 14676–14684. doi:10.1523/JNEUROSCI.3250-10.2010

Zeithamova, D., Schlichting, M. L., & Preston, A. R. (2012). The hippocampus and inferential reasoning: Building memories to navigate future decisions. *Frontiers in Human Neuroscience, 6,* 70. doi:10.3389/fnhum.2012.00070

Appendix

Relationship of REMERGE (Recurrency, and Episodic Memory Results in Generalization) to the Generalized Context Model (GCM)

It is worth noting that the procedure by which the activity of units in the conjunctive layer is calculated is directly analogous to the similarity-based computation (i.e., the similarity of test stimulus *i* to exemplar *j*) described in the exemplar-based GCM model (Nosofsky, 1984). While the distance measure used in GCM is based on the number of mismatching features between test stimulus and stored exemplar, and the $net_i(t)$ term computed by the network is driven by the number of matching features, the equivalence between these two procedures in terms of resulting similarities (or activities) can easily be demonstrated:

According to GCM, the similarity, η, of item *i* to previously experienced exemplar *j* is given by

$$\eta(i, j) = e^{-c*d(i, j)}$$

Where *d* is the raw psychological distance between item *i* and exemplar *j*, and *c* is the sensitivity parameter.

The distance $d(i, j)$ between stimulus *i* and exemplar *j*, omitting attentional weight parameters typically included in GCM, is given by adding up the number of mismatching features—that is, using a city-block metric, under the assumption that the *N* dimensions of a stimulus are separable (e.g., Nosofsky, 1984):

$$d(i, j) = \sum_{k=1}^{N} |x(i, k) - x(j, k)|$$

Where *N* is the number of feature dimensions, and $x(i, k)$ is the value of stimulus *i* on dimension *k*, and $x(j, k)$ is the value of exemplar *j* on dimension *k*.

The similarity of item *i* to previously experienced item *j* can also be expressed in terms of the number of matching features, $m(i, j)$:

$$\eta(i, j) = e^{-c*(N-m(i, j))}$$

or equivalently:

$$\eta(i, j) = e^{-c*N} * e^{c*m(i, j)}$$

with the similarity of item *i* to stored item *j*, relative to all *n* previously experienced exemplars given by

$$S_i = \frac{e^{c*m(i, j)}}{\sum_{j=1}^{n} e^{c*m(i, j)}} \; or \; S_i = \frac{e^{c*m(i, j)}}{k + \sum_{j=1}^{n} e^{c*m(i, j)}}$$

In the latter formulation, *k* denotes a constant term, which captures the probability that the current item is new (i.e., not a previously experienced exemplar). Of note, certain formulations of the GCM have included a related constant term when computing similarity (e.g., see Nosofsky et al., 2012).

GCM's similarity measure, therefore, can be seen to be equivalent to the hedged softmax activation function applied to the conjunctive layer in REMERGE, where the sensitivity parameter, *c*, is replaced by the inverse of the temperature (i.e., $1/\tau$) and the net input to a given conjunctive unit driven by matching features in the input.

(*Appendix continues*)

$$y_i = \frac{e^{net_i/\tau}}{C^{1/\tau} + \sum_{i=1}^{N} e^{net_i/\tau}}$$

As such, it can be appreciated that the sensitivity parameter, c, of GCM is closely related to the temperature parameter, τ, of REMERGE. Lower temperatures have the effect of stretching the psychological space, in effect magnifying the difference between stimuli and, therefore, increasing the emphasis the model places on small differences between stimulus i and exemplar j.

Taken together, the activity of units on the network's conjunctive layer at each timestep can be viewed as reflecting the similarity of respective premise pairs (i.e., stored exemplars) to the current pattern of activity on the feature layer. Recurrence, therefore, is critical in allowing exemplar-based similarity computation to be performed not only on externally presented sensory inputs (i.e., test stimuli displayed on the screen) but also on feature layer inputs reconstructed by the network.

Assessing the Performance of REMERGE in Other Domains: Recognition Memory and Categorization

Here, we consider implications—and, in particular, possible adverse effects—of recurrence in two different domains: namely, categorization and recognition (i.e., episodic) memory. Notably, the hippocampus is widely accepted to play an important role in episodic memory (Brown & Aggleton, 2001; Eichenbaum, Yonelinas, & Ranganath, 2007; Norman & O'Reilly, 2003), and exemplar models have been highly influential in accounting for performance in both recognition memory tasks (Hintzman, 2001; Shiffrin & Steyvers, 1997) and categorization tasks (Medin & Schaffer, 1978; Nosofsky, 1984). As such, we aimed to provide a proof-of-principle demonstration that recurrency is compatible with adequate performance in these settings.

Categorization: The 5–4 Structure

The ability to categorize stimuli into one class or another is a core cognitive function and one that has been the subject of extensive research (Ashby & Maddox, 2005; G. L. Murphy, 2004). Several influential models have been proposed to account for patterns of behavioral performance observed in categorization experiments, and these models differ qualitatively in their views of the nature of category representation. For instance, prototype models argue that categories are represented by the average (i.e., central tendency) of a set of experiences (Ashby & Maddox, 2005). In contrast, exemplar-based models of categorization assume storage of each individual exemplar encountered during training, with new stimuli assigned to the category with the most similar exemplars (Hintzman, 1986; Medin & Schaffer, 1978; Nosofsky, 1984).

Given that exemplar models were developed largely to account for behavioral data in the setting of categorization tasks, it is important to ask whether the REMERGE model, incorporating a notion of recurrency, retains this capacity. Given the expansive nature of the categorization literature, we focus on one particularly popular task, which has been the subject of extensive empirical investigations and where exemplar models have been highly influential, often termed the "5–4 task" (Medin, 1975; Medin & Schaffer, 1978; J. D. Smith & Minda, 2000).

Overview of Experimental Design

While the stimuli themselves are highly variable between experiment (e.g., geometric forms, Brunswick faces), they have four binary dimensions (e.g., size, color, form, number. There are a total of nine training stimuli and seven test stimuli. Five of the training stimuli are derived from the prototype of Category A (1111), and four are derived from the Category B prototype (0000). The composition of category members is detailed in Table A1 and is discussed in detail in J. D. Smith and Minda (2000). Briefly, four exemplars of Category A share three features with the prototype, and one exemplar shares two features (i.e., is ambiguous). Category B has two ambiguous items, one exemplar that shares three features and the prototype itself. Of note, the 5–4 category structure is linearly separable (i.e., can be partitioned by a linear discriminant function). Further, the two categories have little family resemblance and are poorly differentiated (i.e., low structural ratio: exemplars are nearly as similar across categories as within category), an aspect of the task that may lead exemplar-based models to be favored (see J. D. Smith & Minda, 2000).

Table A1
The 5–4 Category Structure

	Dimension (D)			
Stimulus	D1	D2	D3	D4
A1	1	1	1	0
A2	1	0	1	0
A3	1	0	1	1
A4	1	1	0	1
A5	0	1	1	1
B6	1	1	0	0
B7	0	1	1	0
B8	0	0	0	1
B9	0	0	0	0
T10	1	0	0	1
T11	1	0	0	0
T12	1	1	1	1
T13	0	0	1	0
T14	0	1	0	1
T15	0	0	1	1
T16	0	1	0	0

Note. A = Category A; B = Category B; T = Transfer. Adapted with permission from "Thirty Categorization Results in Search of a Model," by J. D. Smith and J. P. Minda, 2000, *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* p. 4. Copyright 2000 by the American Psychological Association.

(*Appendix continues*)

In a typical experiment (see Experiment 2 of Medin, 1975), subjects are first trained on the nine stimuli, comprising five members of Category A and four of Category B. Over multiple training trials, subjects are presented with an exemplar (e.g., two large red triangles), are asked to categorize the stimulus (i.e., as A or B), and receive corrective feedback. In the critical test phase of the experiment, subjects are typically required to assign category labels to the nine training exemplars, together with seven new exemplars (see Table A1), without receiving feedback. Classification probability (i.e., the chance of assigning a given item to Category A during the test phase) forms the typical dependent variable of interest during such experiments.

Model Specifics

The general architecture of the model, consisting of recurrence between featural and conjunctive layers, was tailored to the current setting. Given the four binary dimensions of the exemplars, four stimulus attribute feature pools were included, each consisting of two units. A fifth feature pool, also consisting of two units, was also present, to denote the category label (i.e., A or B) of each exemplar. The conjunctive layer consisted of nine units corresponding to each of the training exemplars.

The following connections were present in the model: bidirectional excitatory connections between the conjunctive layer and the four stimulus attribute feature pools. Unidirectional excitatory connections were present between the conjunctive layer and category feature pool, in line with the distinct status afforded the category denomination in the experiment itself. Of note, similar results were obtained when recurrent connections between conjunctive and category pools were included in the network. A softmax activation function was used over all layers in the network. While in previous simulations (e.g., in the transitive inference task) a logistic activation function was used on the feature layer, here we use a softmax function in line with the mutually exclusive nature of stimulus attributes in the current setting (i.e., square or triangle).

As in previous applications, the network was considered to have stored the nine training examples prior to testing. Testing was implemented in the model by presenting external input to the relevant units in the feature layer. As previously, the network's performance was indexed by the Luce choice ratio, determined by the final activities of units in the category label feature pool. In the current simulations, the amplitude of weights in the network was fixed at 1.

Simulation Results

We first asked whether the categorization performance of REMERGE in this setting follows a similar profile to that of the GCM, a classical example of a non-recurrent exemplar style model. As illustrated in Figure A1, the categorization performance

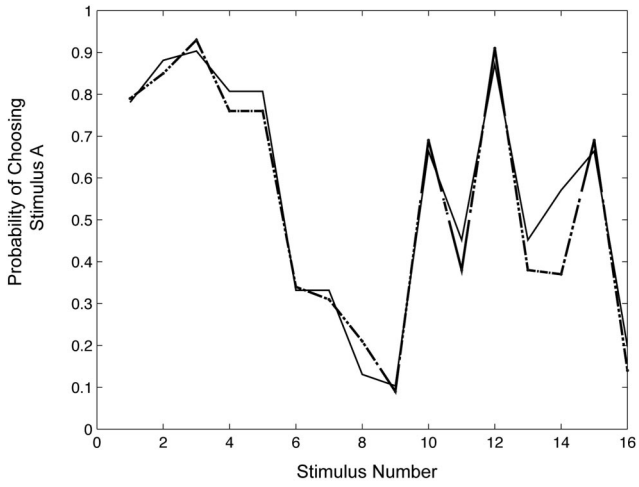

Figure A1. The 5–4 category learning task: probability of assigning each of 16 test stimuli (*x*-axis) to Category A (*y*-axis) according to the generalized context model (GCM; dotted line) and REMERGE (recurrency, and episodic memory results in generalization; solid line). Data relating to GCM are drawn from the intermediate setting of the sensitivity parameter (i.e., 5), described in Nosofsky (2000). Parameters in REMERGE include the following: $\tau = 0.65$; $C = 1$; $\beta = 0.25$. Note that REMERGE also provides an adequate fit to empirical data summarized in a meta-analysis of 30 empirical studies (J. D. Smith & Minda, 2000; though see Nosofsky, 2000, for a discussion of potential issues with this analysis; data not shown: see the main text for details). Also see Table A1 in the Appendix for a description of the 5–4 category structure.

of GCM (Nosofsky, 2000) and REMERGE follows a highly similar profile (sum of squared deviations measure [SSD] = 0.06). In particular, both models exhibit the "A2/A1" advantage (i.e., a higher probability of classifying Item A2 than Item A1 as a member of category A), a prototypical phenomenon shown by exemplar models in this context.

REMERGE also provides a satisfactory fit to empirical data observed in the 5–4 categorization task. J. D. Smith and Minda (2000) reported a meta-analysis of the classification performance of subjects averaged across 30 experiments where this category structure was employed. Though specific concerns (e.g., about the nature of studies included in this meta-analysis) have been raised (Nosofsky, 2000), it is nevertheless worth noting that our model performs relatively well (SSD = 0.07), compared to the spectrum of models analyzed by J. D. Smith and Minda, in fitting the empirical data. Notably, many of these models have more free parameters. For instance, GCM (Nosofsky, 1984) provides a closer fit to this data set (SSD = 0.06) but includes five parameters (plus a stochastic choice parameter): four attentional weighting parameters (for each of the stimulus dimensions) and a sensitivity parameter (directly analogous to the temperature parameter used by the sigmoidal activation functions in our model).

(*Appendix continues*)

These results demonstrate that the REMERGE model appears to retain similar categorization capabilities as traditional exemplar models such as GCM. Further, REMERGE is capable of providing an adequate fit to the empirical data, based on only three free parameters. As such, this simulation confirms that the combination of recurrence and similarity computation does not have inevitable costs for the categorization ability of exemplar models. In the future, it may be useful to develop experimental paradigms involving category structures that might allow the operation of a standard exemplar type mechanism to be teased apart from a recurrent model such as REMERGE.

One point worth noting, however, is that the empirical data leave open the question of whether the hippocampus plays an important role in categorization. While a recent meta-analysis reported that patients with amnesia and MTL damage do indeed show an impairment at categorization (Zaki, 2004; also see Kitchener & Squire, 2000), this conclusion is hotly debated (Knowlton & Squire, 1993; Squire, Stark, & Clark, 2004). Our primary concern, therefore, was to verify that a principle of recurrency is compatible with satisfactory categorization performance, at least relating to one particular structure (i.e., 5–4 structure), rather than necessarily implying that the hippocampal system itself plays an important role in such settings.

Recognition Memory

The ability to judge whether a stimulus, or experience, has been encountered in the past (e.g., recognizing a familiar person when walking down the street) is a critical component of episodic memory and is widely accepted to be critically dependent on neural structures within the medial temporal lobe (Brown & Aggleton, 2001; Norman & O'Reilly, 2003). While it is clear that the hippocampus makes an important contribution to recognition memory, whether this occurs exclusively through the recall of the specific content of prior experiences remains subject to debate (Brown & Aggleton, 2001; Norman & O'Reilly, 2003; Wixted & Squire, 2010). Computational accounts of the hippocampus in recognition memory have highlighted the importance of its unique representational (e.g., pattern separation) and computational (e.g., pattern completion) abilities (e.g., Norman & O'Reilly, 2003). Critically, previous accounts (e.g., Norman & O'Reilly, 2003) have illustrated the recognition memory capacities of the hippocampus in the setting of a traditional unidirectional perspective of the circuit. While exemplar based models of memory have also been applied with success to data from recognition memory tasks (e.g., McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997), it is important to ask whether the incorporation of a recurrent mechanism impairs function in this regard. To address this issue, we compared the recognition memory performance of the REMERGE model with a feedforward version in which recurrence between feature and conjunctive layers was disabled.

Design of Stimuli

As in Norman and O'Reilly (2003), the stimuli were generated from prototypical stimuli. In this case, the two prototypes were anti-correlated with one another (A: 10100110, B: 01011001). The 16 training stimuli consisted of eight distortions of each prototype (4×1 bit distortions, 4×2 bit distortions). Test stimuli consisted of the original 16 training stimuli, as well as eight novel distortions of each prototype (i.e., 4×1 bit, 4×2 bit distortions). New items, therefore, overlapped with their respective prototype in six or seven out of eight features, and with previously studied items from the same category by between four and six features. In a typical recognition memory experiment, a subject would be presented with training stimuli during the study phase of the task and would be asked to give recognition memory judgments during the test phase (i.e., old vs. new decisions).

Model Specifics

Recurrent model. The generic architecture of the model was similar to that used in the categorization simulation described above: Given the eight binary dimensions of the exemplars, eight stimulus attribute feature pools were included, each consisting of two units. The conjunctive layer consisted of 16 units corresponding to each of the training exemplars. The following connections were present in the model: bidirectional excitatory connections between the conjunctive layer and the eight stimulus attribute feature pools. A softmax activation function was used in both the conjunctive and feature pools in the network, in line with the mutually exclusive nature of stimulus attributes in the current setting. As in previous applications, the network was considered to have stored the 16 training examples prior to testing. The amplitude of weights in the network was fixed at 1.0. The temperature parameter was varied to assess recognition performance at different settings.

Feedforward model. Our intention was to use a directly analogous architecture to the recurrent model but with recurrency itself absent. This was implemented in the following way: The feature layer (i.e., consisting of eight two-unit pools) was duplicated. The first feature layer, termed feat_in, projected via unidirectional excitatory connections to the 16-unit conjunctive layer, while the second feature layer, termed feat_out, was in receipt of unidirectional excitatory connections from the conjunctive layer.

Testing Procedure

Testing was implemented in the recurrent model by presenting external input to the relevant units in the feature layer, depending on the test stimulus currently present. To replicate the conditions present in the recurrent model, external input was presented to both feat_in and feat_out in the feedforward model.

To evaluate the recognition performance of each model, we followed the overall procedure employed by Norman and O'Reilly (2003). Specifically, activity over feature units (i.e., over the single feature layer in the recurrent model, and over the feat_out layer in the feedforward model) matching the test stimulus was viewed as

(*Appendix continues*)

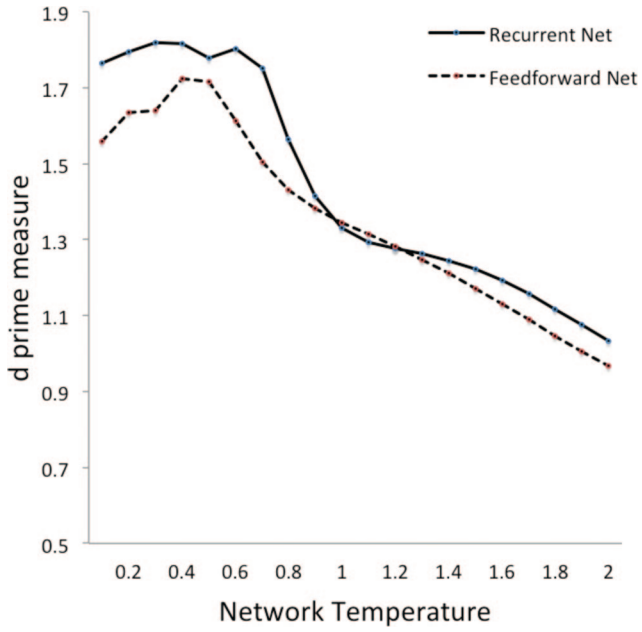

Figure A2. Recognition memory simulation: performance of recurrent and feedforward network, indexed by measure of signal strength (*d*-prime) based on difference in feature layer activity for studied and lure items, shown for network temperatures across the range 0.1 to 2.0 (in increments of 0.1). Note the relatively similar performance of recurrent and feedforward networks, across a relatively large range of network temperatures. See the main text for details.

evidence that the current item had been previously studied. In contrast, activity over feature units that mismatched the current stimulus was used as evidence against the item in question having previously been studied. For each test item, therefore, we calculated a "recall score" in this fashion—that is, the difference between matching and mismatching feature unit activity. For each model, we indexed the strength of the recognition memory signal by using a standard measure of discriminability (*d*-prime)—that is, the difference between the mean amplitudes of recall scores for studied and novel items divided by the mean variance of these recall scores. This procedure was repeated for each model at different network temperatures to assess the effect of this parameter on recognition performance.

The simulation results (see Figure A2) illustrate that the recognition memory performance of the recurrent model is roughly comparable to that of the feedforward model. It can be appreciated that the recognition memory performance of both models is better at lower network temperatures (e.g., 0.1), where typically only the activation of a single (best-matching) conjunctive unit is tolerated. Nevertheless, recognition performance is still maintained at temperatures that allow the co-activation of multiple conjunctive units (e.g., 1.0), in both recurrent and feedforward models. As such, it is important to note that the satisfactory recognition performance is achieved within a relatively large range of network temperatures.

This simulation, therefore, provides a proof-of-principle demonstration that the addition of a recurrent mechanism to an exemplar style model does not have inevitable costs for recognition memory performance—these results, therefore, support the notion that recurrency in the hippocampal circuit is broadly compatible with its well-established role in episodic memory.

Received August 8, 2010
Revision received April 10, 2012
Accepted April 19, 2012 ■