Available online at www.sciencedirect.com
SCIENCE DIRECT
Neural Networks 18 (2005) 1212–1228
Neural Networks
www.elsevier.com/locate/neunet

2005 Special issue

Methods for reducing interference in the Complementary Learning Systems model: Oscillating inhibition and autonomous memory rehearsal

Kenneth A. Norman *, Ehren L. Newman, Adler J. Perotte

Department of Psychology Princeton University, Green Hall, Princeton, NJ 08544, USA

Abstract

The stability–plasticity problem (i.e. how the brain incorporates new information into its model of the world, while at the same time preserving existing knowledge) has been at the forefront of computational memory research for several decades. In this paper, we critically evaluate how well the Complementary Learning Systems theory of hippocampo–cortical interactions addresses the stability–plasticity problem. We identify two major challenges for the model: Finding a learning algorithm for cortex and hippocampus that enacts selective strengthening of weak memories, and selective punishment of competing memories; and preventing catastrophic forgetting in the case of non-stationary environments (i.e. when items are temporarily removed from the training set). We then discuss potential solutions to these problems: First, we describe a recently developed learning algorithm that leverages neural oscillations to find weak parts of memories (so they can be strengthened) and strong competitors (so they can be punished), and we show how this algorithm outperforms other learning algorithms (CPCA Hebbian learning and Leabra at memorizing overlapping patterns. Second, we describe how autonomous re-activation of memories (separately in cortex and hippocampus) during REM sleep, coupled with the oscillating learning algorithm, can reduce the rate of forgetting of input patterns that are no longer present in the environment. We then present a simple demonstration of how this process can prevent catastrophic interference in an AB–AC learning paradigm.

Keywords: Hippocampus; Neocortex; Neural network; Interference; Learning algorithm; Theta Oscillations; Sleep; Consolidation

1. Introduction

1.1. Stability–plasticity

Over the past several decades, neural theorists have converged on the idea that neocortex implements an internal, predictive model of the structure of the environment. This internal model must simultaneously maintain previously learned information and integrate new information. The problem of how to accomplish these goals simultaneously in a neural network architecture was labeled the *stability–plasticity dilemma* by Carpenter and Grossberg (1988), and this problem has come to occupy a central position in computational neuroscience. The problem is hard to solve because, in most neural network models, memory traces overlap with one another. As such, learning new memories will incrementally degrade pre-existing memories. Several researchers have found that, when new learning is extensive (e.g. if the system has to memorize a new pattern based on a single learning trial), neural networks can show near-complete forgetting of pre-existing knowledge (*catastrophic interference*; French, 1999; French, 2003; McCloskey & Cohen, 1989).

There have been several attempts to solve this problem, e.g. Adaptive Resonance Theory (Carpenter & Grossberg, 2003]). In this paper, we focus on another framework for addressing stability–plasticity: The Complementary Learning Systems (CLS) model (McClelland, McNaughton, & O'Reilly, 1995; O'Reilly & Norman, 2002; O'Reilly & Rudy, 2001; Norman & O'Reilly, 2003). This model posits that cortex solves stability–plasticity with the assistance of a hippocampal system that can rapidly memorize events and play them back to cortex in an 'off-line' fashion. In Section 1.2, we describe the basic properties of CLS, and how it is meant to solve stability–plasticity.

We also briefly review some of the many ways in which CLS has been applied to episodic memory and animal learning data. However, while CLS has proved to be a very useful way for thinking about hippocampal and cortical learning processes, in recent years we have identified some issues with the model that we want to address:

* Corresponding author.
E-mail addresses: knorman@princeton.edu (K.A. Norman), enewman@princeton.edu (E.L. Newman), ajp2120@columbia.edu (A.J. Perotte).

0893-6080/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neunet.2005.08.010

- The first issue involves finding a suitable algorithm for adjusting synapses in cortex and the hippocampus. Some of the learning algorithms that have been used in CLS implementations (e.g. CPCA Hebbian learning: Norman & O'Reilly, 2003; O'Reilly & Munakata, 2000) adjust synapses more than is necessary and, as such, show unacceptably high levels of interference. Other learning rules that have been used in CLS implementations (e.g. *Leabra*; O'Reilly & Munakata, 2000) are less prone to this problem, but have other problems of their own (e.g. both Leabra and CPCA Hebbian learning have difficulty in modeling data on how competitors are punished during retrieval).
- The second issue involves the problem of non-stationary environments: What happens when patterns that were originally in the training set are removed from the training set? Even with the hippocampus and cortex working together, the standard form of the CLS model shows unacceptably high rates of forgetting of patterns once they are removed from the training set. This problem needs to be addressed before the CLS model can be viewed as a complete solution to the stability–plasticity problem.

In this paper, we present solutions to both of these problems:

- In section 2, we describe a new learning algorithm developed by Norman, Newman, Detre, and Polyn (2005) that leverages regular oscillations in feedback inhibition to pinpoint weak parts of target memories (so they can be strengthened) and to pinpoint non-target memories that compete with target memories during retrieval (so they can be weakened). We show that the oscillating learning algorithm, applied to our cortical network, outperforms both CPCA Hebbian learning and Leabra on a pattern completion task. We also show that the oscillating learning algorithm's capacity for supporting familiarity discrimination greatly exceeds the capacity of the Hebbian cortical model from Norman and O'Reilly (2003).
- In section 3, we show how the CLS model can be supplemented by a new kind of off-line learning where cortex and hippocampus separately rehearse stored memories, thereby repairing damage to these memories. We argue that this off-line learning reflects the functionality of REM sleep, and show that it can successfully prevent loss of knowledge in an AB–AC interference paradigm (where AB items are initially trained and then removed from the training set).

In summary: We will present an account of how inhibitory oscillations and off-line rehearsal of stored knowledge (during REM sleep) can both improve learning and retention. The ideas presented here apply to both hippocampus and cortex. For simplicity's sake, the simulations that we present will use the cortical model, which has a less differentiated architecture than the hippocampal model. After each simulation, we will discuss ways in which the same mechanism can be applied to the hippocampus.

1.2. Basic properties of CLS

The CLS framework (McClelland et al., 1995) incorporates several widely-held ideas about hippocampal and neocortical contributions to memory, that have been developed over many years by many different researchers (e.g. Aggleton & Brown, 1999; Burgess & O'Keefe, 1996; Eichenbaum, Otto, & Cohen, 1994; Grossberg, 1976; Hasselmo & Wyble, 1997; Marr, 1971; McNaughton & Morris, 1987; Moll & Miikkulainen, 1997; O'Keefe & Nadel, 1978; Rolls, 1989; Scoville & Milner, 1957; Sherry & Schacter, 1987; Squire, 1992; Sutherland & Rudy, 1989; Teyler & Discenna, 1986; Treves & Rolls, 1994; Wu, Baxter, & Levy, 1996; Yonelinas, 2002). According to the CLS framework, neocortex forms the substrate of our internal model of the structure of the environment. In contrast, hippocampus is specialized for rapidly and automatically memorizing patterns of cortical activity, so they can be recalled later (based on partial cues).

The CLS framework posits that neocortex learns incrementally; each training trial results in relatively small adaptive changes in synaptic weights. These small changes allow cortex to gradually adjust its internal model of the environment in response to new information. The other key property of neocortical learning is that it assigns similar (overlapping) representations to similar stimuli. Use of overlapping representations allows cortex to represent the shared structure of events, and therefore makes it possible for cortex to generalize to novel stimuli based on their similarity to previously experienced stimuli. In contrast, hippocampus is biased to assign distinct, *pattern separated* representations to stimuli, regardless of their similarity. This property allows hippocampus to rapidly memorize arbitrary patterns of cortical activity without suffering catastrophic levels of interference.

1.3. How CLS solves stability–plasticity

One of the key problems facing any account of stability–plasticity is how to incorporate rare (but significant) events into the cortical network. In the case of the CLS model, the incremental nature of cortical learning means that it can only retrieve memories if the stimulus is presented repeatedly.

However, infrequently-occurring events are sometimes very significant (e.g. if a pterodactyl eats your sister) and we need to be able to incorporate this information into our internal cortical model of how the world works, so we can properly generalize to new situations (e.g. future pterodactyl attacks). If the cortical network were left to its own devices, a person would have to experience several pterodactyl attacks before the cortical memory trace was strong enough to support appropriate recall and generalization. Furthermore, if the average interval between pterodactyl appearances were sufficiently long, one runs the risk that—in between appearances—interference from other memories would erode the original memory, in which

case the person would be back to where they started with each new pterodactyl appearance.

The presence of the hippocampal network solves this problem. The hippocampus is specialized for rapid memorization; in a single trial, the hippocampus can latch on to pattern of cortical activity elicited by the pterodactyl, and re-play it to cortex repeatedly until it sinks in. In this respect, hippocampus can be viewed as a 'training trial multiplier'. Over time, hippocampally-mediated replay of pterodactyl memories is interleaved with bottom-up learning about information in the environment. As discussed by McClelland et al. (1995), this kind of interleaved training, coupled with a learning mechanism that is sensitive to prediction error, forces cortex to develop representations that reconcile the properties of rare events and more common events (because this is the only way to avoid prediction error across the entire training set).[1]

1.4. Applications of CLS to episodic memory and other domains

CLS was originally formulated as a set of high-level principles for understanding hippocampal and cortical contributions to memory. More recently, O'Reilly and Rudy (2001) and Norman and O'Reilly (2003) have developed working neural network models of hippocampus and neocortex that instantiate these principles, and these networks have been applied to modeling specific datasets.

1.5. Modeling recognition memory

In one application, Norman and O'Reilly (2003) implemented hippocampal and cortical networks that adhere to CLS principles, and showed how these networks (taken together) constitute a *neural dual-process model* of recognition memory. Learning was implemented in these simulations using a simple Hebbian rule (called *instar learning* by Grossberg, 1976, and *CPCA Hebbian learning* by O'Reilly & Munakata, 2000), whereby connections between active sending and receiving neurons are strengthened, and connections between active receiving neurons and inactive sending neurons are weakened. Norman and O'Reilly (2003) showed how the hippocampal model (using this simple Hebbian rule) can support recognition via recollection of specific studied details. The cortical model cannot support recollection of specific details from once-presented events, owing to its relatively low learning rate. However, Norman and O'Reilly (2003) showed that cortex can still support judgments of familiarity after a single study trial, based on the *sharpness* of representations in cortex.

The cortical model's ability to support familiarity discrimination is a simple consequence of Hebbian learning and inhibitory competition. When a stimulus is presented, Hebbian learning tunes a subset of the hidden units to respond more strongly to that stimulus. As these units respond more and more strongly to the stimulus, they start to inhibit other units. Thus, the neural response to a stimulus transitions from a diffuse overall response (where no units are tuned to respond strongly to the stimulus) to a more focused response where some units are strongly active and other units are suppressed. In the Norman and O'Reilly (2003) paper, cortical familiarity was operationalized in terms of the activation of the k most active units in the hidden layer (where k is a model parameter that defines the maximum number of units that are allowed to be strongly active at once), although other methods of operationalizing familiarity are possible.

Norman and O'Reilly (2003) showed how, taken together, the hippocampal network and cortical network could explain a wide range of recognition findings, including data on when hippocampal lesions affect recognition memory (as a function of how similar distractors are to studied items, and as a function of test format) and data from normal subjects on how interference manipulations affect recognition memory (e.g. list strength manipulations: how does repeatedly presenting some items on the study list affect memory for other items on the study list).

1.6. Modeling animal learning

In another application, O'Reilly and Rudy (2001) used hippocampal and cortical networks instantiating CLS principles to explain findings from animal learning paradigms, including non-linear discrimination learning (e.g. negative patterning, transverse patterning), 'transitive inference' in discrimination learning, and contextual fear conditioning. The models in these simulations were largely identical to the models used in Norman and O'Reilly (2003), except the simulations used O'Reilly's Leabra learning rule instead of CPCA Hebbian learning. Leabra combines CPCA Hebbian learning with a simple form of error-driven learning (O'Reilly & Munakata, 2000). The key finding from these simulations was that cortex could solve non-linear discrimination problems on its own when the animal is given repeated exposure to the stimuli and appropriate feedback. In contrast, hippocampus is needed to show sensitivity to feature conjunctions on tasks where conjunctive learning is incidental (i.e. the animal does not have to learn the conjunction to respond correctly on the task) and the animal is given limited exposure to the conjunction. O'Reilly and Rudy (2001) discuss several findings that support the model's predictions.

1.7. Problems with learning rules

Concrete applications of CLS (like those described in Norman & O'Reilly, 2003 and O'Reilly & Rudy, 2001) have

[1] One could reasonably ask why we need to represent rare events in cortex, given that hippocampus is capable of recalling these events after a single training trial. The answer (according to CLS) is that, even though hippocampus can support recall, it is not well suited to feature-based generalization. Thus, to the extent that we want to generalize properly to similar events in the future (e.g. different colors and sizes of pterodactyls appearing in different locations), information about pterodactyls needs to be transferred from hippocampus to cortex.

provided strong support for the validity of basic CLS principles (see also O'Reilly & Norman, 2002). However, the process of building working models that instantiate CLS principles has also highlighted some important challenges for the CLS framework.

One critical challenge is to develop a learning algorithm that is capable of storing an appropriately large database of knowledge (semantic knowledge, in the case of cortex, and episodic knowledge, in the case of the hippocampus). Norman and O'Reilly (2003) noted that the CPCA Hebbian learning rule used in that paper has a tendency to over-focus on prototypical features. When given a large set of correlated input patterns to memorize, the CPCA Hebbian algorithm is very good at learning what all of these patterns have in common, but it shows very poor memory for specific, non-prototypical features of individual items. This is less of a problem for the hippocampal model than for the cortical model, because of the hippocampal model's ability to assign relatively distinct representations to similar inputs. However, Norman and O'Reilly (2003) noted that the hippocampal model is still prone to 'pattern separation collapse' when given large numbers of overlapping patterns. When this occurs, the hippocampus recalls prototypical features in response to all input patterns (studied or non-studied).

From a psychological-modeling perspective, the mere fact that Hebbian learning over-focuses on prototypes is not problematic. Good memory for prototypes can be used to explain numerous categorization and memory phenomena (e.g. false recognition of non-studied items from studied categories; Koutstaal, Schacter, & Jackson, 1999). Also, as discussed by Norman and O'Reilly (2003), the model's tendency to forget individuating features of studied items can be used to explain memory interference effects on list learning paradigms.

However, the excessive degree of prototype-focusing exhibited by the model is more problematic. When the model is given a sufficiently large number of overlapping patterns, both the hippocampal and cortical networks exhibit virtually no memory for individuating features. In an important analysis, Bogacz and Brown (2003) set out to quantify the capacity of several different cortical models (including the Norman & O'Reilly, 2003 Hebbian cortical network) for supporting familiarity-based recognition: How many patterns can be stored in the network, in a manner that supports discrimination of studied vs. non-studied patterns? This analysis showed that, given overlapping input patterns, the capacity of the Hebbian cortical network from Norman and O'Reilly (2003) was very poor. Even in a brain-sized version of the network, the model's capacity is almost certainly not large enough to account for data on human recognition memory capacity (e.g. Standing, 1973) showed that people can discriminate between thousands of studied vs. non-studied pictures, and this is an extremely conservative estimate).

1.8. Why does CPCA Hebbian learning perform poorly?

The essence of the problem with CPCA Hebbian learning is that it is insufficiently judicious in how it adjusts synaptic strengths. In neural networks, each synaptic weight is involved in storing multiple memories. As such, adjusting weights to improve recall of one memory interferes with other memories that are encoded in those weights. Given that there is a cost (in terms of interference) as well as a benefit to adjusting synaptic weights, it makes sense that strengthening of weights should stop once the target memory is strong enough to support recall and generalization. Likewise, learning algorithms should only weaken non-target memories that are actively competing with recall of the target memory. Any further strengthening (of the target memory) or weakening (of non-target memories) will cause interference without improving recall. CPCA Hebbian learning fails on both counts: It strengthens synapses between co-active units even if the target memory is already strong enough to support recall, and it weakens synapses between active receiving units and all sending units that are inactive at the end of the trial, even if these units did not actively compete with recall of the target memory.

In addition to being inefficient (from a functional standpoint), CPCA Hebbian learning's inability to selectively weaken competing memories also impedes its ability to account for empirical data on competitor punishment. Over the past decade, several studies have found that memory weakening is modulated by how strongly memories compete at retrieval: Non-target memories that compete strongly with the target memory (but subsequently lose the competition to be retrieved) are punished. However, if steps are taken to mitigate competition (e.g. by increasing the specificity of the retrieval cue), there is no punishment (see Anderson, 2003 for a review of these findings; see also Norman, Newman, & Detre, 2004 for a computational model of these findings). This pattern of results has been observed in both semantic memory tasks (e.g. Blaxton & Neely, 1983) and episodic memory tasks (e.g. Anderson & Bell, 2001; Ciranni & Shimamura, 1999), suggesting that selective competitor punishment occurs in both cortex and hippocampus. However, contrary to these findings, CPCA Hebbian learning predicts that all memories that overlap with the target memory should be weakened, regardless of the amount of competition at retrieval.

1.9. Problems with Leabra

As mentioned earlier, some implementations of CLS (e.g. O'Reilly & Rudy, 2001) have used O'Reilly's Leabra learning algorithm instead of CPCA Hebbian learning. Because of its ability to learn based on pattern completion error, Leabra does a much better job than CPCA Hebbian learning at retaining the individuating features of studied items. However, as discussed in Norman et al. (2004), Leabra lacks a mechanism for selectively punishing memories that compete at retrieval. The essence of this problem is that competitor activity is transient (i.e. the competitor 'pops up' briefly and then goes away), but Leabra is only equipped for learning about representations that are active in the final settled state of the network. As such, Leabra also fails to account for the competitor-punishment data discussed above.

1.10. Desiderata for a replacement algorithm

Because of the issues with CPCA Hebbian learning and Leabra outlined above, we set out to derive a new learning algorithm that meets the following two desiderata:

- *Limits on strengthening:* The network should only strengthen memories when they are too weak to support recall.
- *Targeted punishment:* The network should only weaken memories when they actively compete with successful recall of the target memory.

These properties, taken together, should reduce interference in the cortical and hippocampal models. The second property should help the networks account for data on competitor punishment.

2. The oscillating learning algorithm

To meet the desiderata outlined above, Norman et al. (2005) developed a new learning algorithm that selectively strengthens weak parts of target memories (vs. parts that are already strong), and selectively punishes strong competitors. The learning algorithm accomplishes this goal by oscillating the strength of feedback inhibition, and learning based on the resulting changes in activation. In this section, we first provide some background information on how inhibition was implemented in the model, and how the network was structured. We then provide a highlevel overview of how the algorithm works. Finally, we present benchmark data (taken in part from Norman et al., 2005) comparing the oscillating learning algorithm to Leabra and CPCA Hebbian learning.

2.1. Background: how inhibition was implemented in the model

In the simulations described below, we used the simple two-layer cortical network shown in Fig. 1. The network was provided with patterns to memorize on the input/output layer, and the hidden layer was free to self-organize. Every input/output unit was connected to every input/output unit (including itself) and to every hidden unit. Whenever a network is recurrently connected in this manner, there has to be some mechanism for limiting the spread of excitatory activity. In the brain, this problem is solved by inhibitory interneurons, which enforce a *set point* on the amount of excitatory activity within a subregion (O'Reilly & Munakata, 2000). We capture this set point dynamic in our model using a *k-winners-take-all* (*kWTA*) inhibition rule, which adjusts inhibition such that the k units in each layer that receive the most excitatory input are strongly active, and all other units are at most weakly active (activity $<$ .25; Minai & Levy, 1994; O'Reilly & Munakata, 2000). We set the input/output layer k equal to the number of units in each studied pattern, such that (when kWTA is applied to the network) the best-fitting memory—and only that memory—is active.

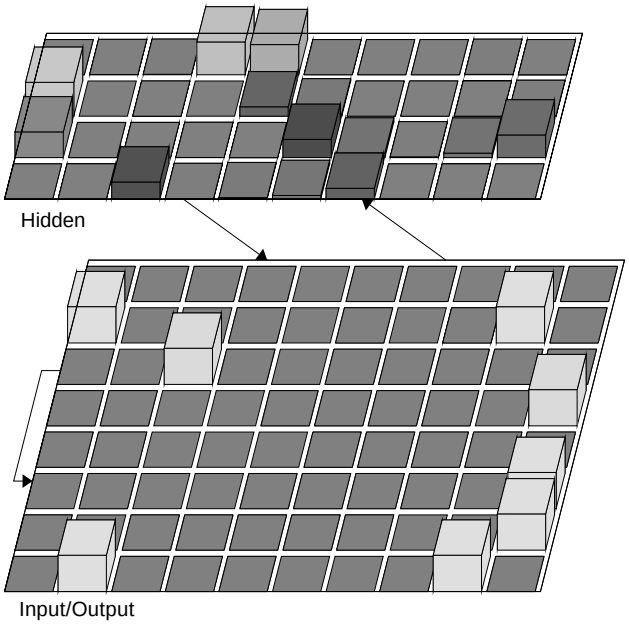

Fig. 1. Diagram of the network used in our simulations. Patterns were presented to the lower part of the network (the *input/output* layer). The upper part of the network (the *hidden* layer) was allowed to self-organize. Every unit in the input/output layer was connected to every input/output unit (including itself) and to every hidden unit via modifiable, symmetric weights.

2.2. Algorithm overview

The learning algorithm can be described in the following five steps:

- First, the target pattern is soft-clamped onto the input/output layer of the network. This soft-clamp is applied for the duration of the trial. Given normal levels of inhibition, the kWTA rule prevents activation from spreading to other units in the input/output layer.
- Second, the algorithm identifies competitors by lowering inhibition below the level specified by kWTA. Effectively, lowering inhibition lowers the threshold amount of excitation needed for a unit to become active. If a non-target unit is just below threshold (i.e. it is receiving strong input, but not quite enough to become active) lowering inhibition will cause that unit to become active.
- Third, the algorithm weakens units that turn on when inhibition is lowered (i.e. strong competitors) by reducing weights from other active units. By doing this, the learning algorithm ensures that a unit that competes on one trial will receive less input the next time that cue is presented. If the same cue is presented repeatedly, eventually the input to that unit will diminish to the point where it no longer activates in the low inhibition condition (so no further punishment occurs). Norman et al. (2004) describe how this property allows the model to simulate detailed patterns of behavioral data on competitor-punishment.
- Fourth, the algorithm identifies weak parts of target memories by raising inhibition above the level specified by kWTA. If a target unit is receiving relatively little

collateral support from other target units, such that its net input is just above threshold, raising inhibition will trigger a decrease in the activation of that unit.

- Fifth, the algorithm strengthens units that turn off when inhibition is raised (i.e. weak target units) by increasing weights from other active units. By doing this, the learning algorithm ensures that a target unit that drops out on a given trial will receive more input the next time that cue is presented. If the same pattern is presented repeatedly, eventually the input to that unit will increase to the point where it no longer drops out in the high inhibition condition (so no further strengthening occurs).

2.3. Algorithm details

The algorithm uses Contrastive Hebbian Learning (CHL; Ackley, Hinton, & Sejnowski, 1985; Hinton, 1989; Hinton & McClelland, 1988; Hinton & Sejnowski, 1986; Movellan, 1990) to enact the weight changes described above. CHL involves contrasting a more desirable state of network activity (the *plus* state) with a less desirable state of network activity (the *minus* state). The CHL equation adjusts network weights so that the more desirable state of network activity is more likely to occur in the future. The following equation shows how weight changes are computed by CHL:

$$dW_{ij} = \varepsilon((X_i^+ Y_j^+) - (X_i^- Y_j^-)) \quad (1)$$

In the above equation, X_i is the activation of the presynaptic (sending) unit, Y_j is the activation of the postsynaptic (receiving) unit. The '+' and '−' superscripts refer to plus-state and minus-state activity, respectively. dW_{ij} is the change in weight between the sending and receiving units, and ε is the learning rate parameter.

The oscillating learning algorithm generates minus states by varying inhibition around the level set by kWTA. When inhibition is at its normal level (i.e. the level set by kWTA), all of the target units (and only those units) will be active. This is the maximally correct state of network activity. On each trial, we distort this pattern by oscillating inhibition in a continuous fashion from its normal level to lower-than-normal, then to higher-than-normal, then back to normal, and we apply the CHL equation to successive time steps of network activity. At each point in the inhibitory oscillation, inhibition is either moving toward its normal level (the 'maximally correct' state), or it is moving away from this state. If inhibition is moving toward its normal level, then the activity pattern at time $t+1$ will be more correct than the activity pattern at time t. In this situation, we use the CHL equation to adapt weights to make the pattern of activity at time t more like the pattern at time $t+1$. However, if inhibition is moving away from its normal level, then the activity pattern at time $t+1$ will be less correct than the activity pattern at time t (it will either contain too much or too little activity, relative to the target pattern). In this situation, we use the CHL equation to adapt weights to make the pattern of activity at time $t+1$ more like the pattern at time t. These rules are formalized in Eqs. (2) and (3).

If inhibition is returning to its normal value:

$$dW_{ij} = \varepsilon((X_i(t+1)Y_j(t+1) - (X_i(t)Y_j(t))) \quad (2)$$

If inhibition is moving away from its normal value:

$$dW_{ij} = \varepsilon((X_i(t)Y_j(t) - (X_i(t+1)Y_j(t+1))) \quad (3)$$

For a detailed description of how the algorithm was implemented, see Norman et al. (2005).

2.4. Relation to neural oscillations

Although the algorithm was not specifically developed as a theory of neural oscillations, it nonetheless may help to explain how neural oscillations contribute to learning. In particular, theta oscillations (rhythmic changes in local field potential at a frequency of approximately 4–8 Hz in humans) have several properties that resonate with the learning algorithm proposed here:

- Theta oscillations depend critically on changes in the firing of inhibitory interneurons (Buzsaki, 2002; Toth, Freund, & Miles, 1997).
- Theta oscillations have been observed in both of the major CLS structures (cortex and hippocampus; for a review, see Kahana, Seelig, & Madsen, 2001).
- Theta oscillations are fast enough to support several complete oscillations per stimulus presentation, and slow enough to allow competitors to activate when inhibition is lowered.
- Theta oscillations have been linked to learning, in both animal and human studies (e.g. Raghavachari et al., 2001). Several studies have found that the direction of potentiation (LTP vs. LTD) depends on the phase of theta (peak vs. trough; Holscher, Anwyl, & Rowan, 1997; Huerta & Lisman, 1996; Hyman, Wyble, Goyal, Rossi, & Hasselmo, 2003). This result mirrors the property of our model whereby the high-inhibition phase of the oscillation is primarily concerned with strengthening target memories (LTP) and the low-inhibition phase of the oscillation is primarily concerned with weakening competitors (LTD).

Given these facts, it seems possible to us that theta oscillations may serve as the neural substrate of the algorithm described here (Norman et al., 2005). However, at this point the linkage is only suggestive, and needs to be confirmed through further investigation.

2.5. Pattern completion simulations

To explore the oscillating algorithm's ability to avoid pattern separation failure and recall individuating features, Norman et al. (2005) ran simulations comparing pattern completion performance for the oscillating algorithm vs. Leabra. In one set of simulations, Norman et al. (2005) gave the network 200 binary input patterns to learn, with 57%

average overlap between patterns. The network was repeatedly presented with the 200-pattern set until learning reached asymptote. At the end of training, pattern completion was tested by measuring the network's ability to recall a single, non-prototypical feature from each pattern, given all of the other features of that pattern as a retrieval cue.

Norman et al. (2005) were also interested in comparing the robustness of the representations learned by each algorithm: To what extent can these representations support retrieval when test cues do not exactly match studied patterns? To get at this issue, they distorted retrieval cues (by adding Gaussian noise to the input pattern that was clamped onto the network) and measured how pattern completion performance varied as a function of the amount of test-pattern noise.

Fig. 2 shows the number of patterns (out of 200) successfully recalled at the end of training by the oscillating learning algorithm and Leabra, as a function of the amount of noise applied to retrieval cues at test (Norman et al., 2005). For comparison purposes, we have also included the results of new simulations using CPCA Hebbian learning. In keeping with the idea (stated earlier) that CPCA Hebbian has difficulty in learning about non-prototypical features, this algorithm performed very poorly, even for low levels of noise. Leabra performed better than CPCA Hebb; this is because the error-driven component of Leabra explicitly computes pattern completion error at training, and adjusts weights to reduce this error. When test noise was set to zero, Leabra and the oscillating algorithm performed comparably. However, when the models were given noisy test cues, the oscillating algorithm performed much better than Leabra.

The oscillating learning algorithm outperforms Leabra in this situation because the oscillating algorithm does a better job

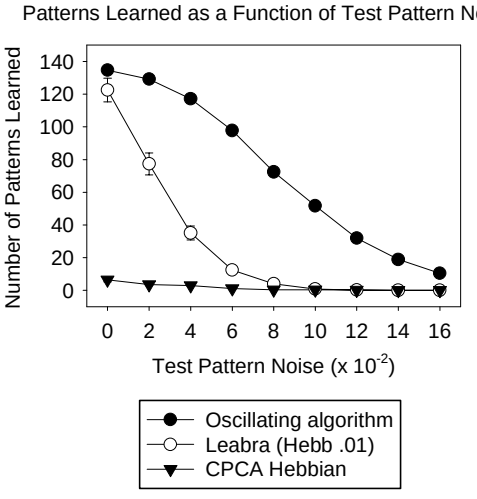

Fig. 2. Comparison of pattern completion performance for the oscillating learning algorithm vs. other learning algorithms. The figure shows the number of patterns (out of 200) successfully recalled at the end of training by each algorithm, as a function of the amount of noise applied to retrieval cues at test; the oscillating-algorithm and Leabra results are taken from Norman et al. (2005). CPCA Hebbian learning performs very poorly. The oscillating learning algorithm and Leabra perform comparably for low noise values, but the oscillating algorithm performs much better than Leabra for noisy retrieval cues.

of maintaining pattern separation in the hidden layer: At the end of training, the average pairwise similarity between patterns in the hidden layer (measured using cosine) was .47 (SEM = .02) for the oscillating algorithm vs. .65 (SEM = .01) for Leabra. The high level of hidden-layer overlap in the Leabra simulations hurts recall by increasing the odds that (given a noisy input pattern) the network will slip out of the correct attractor into a neighboring attractor. The oscillating learning algorithm manages to avoid these pattern-separation difficulties because of its ability to punish competitors: Whenever memories start to blend together, they also start to compete with one another at retrieval, and the competitor-punishment mechanism pushes them apart.

Crucially, even though pattern separation is higher for the oscillating learning algorithm vs. Leabra, the oscillating algorithm still learns similarity-based representations (i.e. it assigns similar hidden representations to similar inputs). The quantify this tendency, we computed a 'similarity score' that tracks the correlation (across all pairs of patterns) between input-layer similarity and hidden-layer similarity. The average similarity score for the oscillating algorithm was .58 (SEM = .02), vs. .71 (SEM = .04) for Leabra. Although the mean similarity score for Leabra was higher, similarity scores for Leabra were also much more variable: Across runs, some scores were extremely high, and some scores were extremely low. Approximately 10% of the Leabra similarity scores were less than .1, indicating a near-total failure to represent the structure of the input space. In contrast, only two runs of the oscillating-algorithm model yielded similarity scores below .5, and these scores (.35 and .45) still showed substantial sensitivity to the structure of the input space.

Finally, the fact that oscillating algorithm learns similarity-based representations (given a cortical network architecture) highlights an essential difference between the pattern separation mechanisms that are built into the CLS hippocampal model, and the pattern separation enacted by the oscillating algorithm. As discussed earlier, the goal of hippocampal pattern separation is to assign maximally distinct representations to stimuli, regardless of their similarity, so these stimuli can be recalled in detail. The hippocampal model's extreme approach to pattern separation effectively cripples its ability to generalize. In contrast, the oscillating algorithm is only concerned that memories observe a 'minimum separation' from one another. So long as this constraint is met, memories in the cortical network simulated here are free to overlap according to their similarity (thereby allowing the network to enact similarity-based generalization).

2.6. Familiarity discrimination: comparison with Hebbian model

In addition to the pattern completion simulations described in Norman et al. (2005), we have recently started to use the oscillating algorithm to simulate familiarity-based recognition in the cortical network. It is possible to read out a familiarity score from the oscillating algorithm by looking at how activation changes when inhibition is raised above its baseline

value: Weak (unfamiliar) memories show a larger decrease in activation than strong (familiar) memories.

In new simulation work (not published elsewhere), we have found that the capacity of the oscillating algorithm for supporting familiarity discrimination is much higher than the capacity of the Hebbian familiarity discrimination model used by Norman and O'Reilly (2003). For example, in one simulation we generated 200 patterns with 41% average overlap. We trained the network by presenting 100 of the patterns for 10 epochs. After each epoch of training, we tested the network's ability to discriminate between the 100 patterns it studied, and the 100 patterns that it did not study. For the oscillating-algorithm familiarity simulations, we used the same network that we used in the pattern completion simulations above (with 80 input units and 40 hidden units). We compared the results of the oscillating-algorithm simulations to the results of simulations using the feedforward Hebbian model from Norman and O'Reilly (2003). The only change to the Hebbian model as described in that paper is that we used 80 input units instead of 240. The exact same input patterns were presented to the oscillating-algorithm model and the Hebbian model. The Hebbian model simulations operationalized familiarity using the *activation of winners* Familiarity measure that was introduced by Norman and O'Reilly (2003): familiarity is the average activation of the k most active hidden units (where k is the activation limit imposed by the k-winners-take-all inhibition rule). For the oscillating-algorithm simulations, we used two different familiarity measures. In one set of simulations, we indexed familiarity in terms of the change in average activation (over the entire input layer) given high vs. normal inhibition. Also, to maximize comparability with the Hebbian simulations, we ran another set of oscillating-algorithm simulations using the activation of winners familiarity measure from Norman and O'Reilly (2003).

The results of these simulations are shown in Fig. 3. The Hebbian model performs just above chance after one epoch of training, and actually gets slightly worse with additional training. This result is robust to a wide range of parameter settings, including hidden layer size—it appears that there is simply no way to get the Hebbian model to show good discrimination of patterns with this level of overlap. This is because of the Hebbian model's tendency to over-focus on prototype features. In contrast, after 10 epochs, the oscillating learning algorithm showed $>92\%$ accuracy in discriminating between studied and non-studied patterns using the same familiarity measure (activation of winners) that was used in the Hebbian model. Asymptotic accuracy was even better ($>99\%$) when we used a familiarity measure (change in activation, given high vs. normal inhibition) that was specifically tailored to the oscillating algorithm. Although we have not yet carried out the requisite mathematical analyses, we think it is quite possible that the oscillating algorithm's capacity for supporting familiarity-based discrimination, in a brain-sized network, will be large enough to account for the vast capacity of human familiarity discrimination (as illustrated, e.g. by Standing, 1973).

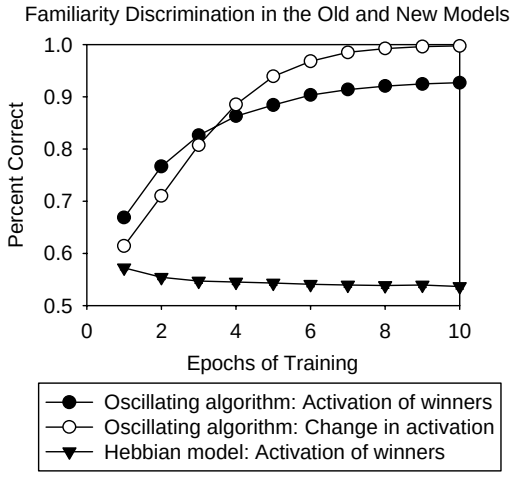

Fig. 3. Comparison of familiarity discrimination using the oscillating learning algorithm vs. the Norman and O'Reilly (2003) Hebbian cortical familiarity model. For the Hebbian model, familiarity was operationalized using the *activation of winners* measure from Norman and O' Reilly (2003). For the oscillating-algorithm model, familiarity was operationalized in two different ways: activation of winners, and also the *change in activation* given high vs. normal inhibition. Note that the oscillating-algorithm simulations used 40 hidden units, whereas the Hebbian simulations used 1920 hidden units (to match the simulations from Norman & O'Reilly, 2003). Despite this large disparity in hidden layer size, the oscillating-algorithm familiarity model strongly outperformed the Hebbian model: Given 100 patterns (and 41% average overlap between patterns), the asymptotic accuracy of the oscillating-algorithm simulations was $>99\%$ for the change in activation measure and $>92\%$ for the activation of winners measure, whereas the Hebbian model's asymptotic accuracy was close to chance.

2.7. Extending the oscillating algorithm to the hippocampal model

The basic principles of the oscillating algorithm (regarding how changes in the strength of inhibition can be used to identify weak parts of target memories, and to flush out competitors) should apply to the hippocampus just as well as they apply to cortex. However, as discussed by Norman et al. (2005), our ideas regarding the functional role of theta oscillations differ from other published theories of how theta contributes to hippocampal processing. Most prominently, Hasselmo, Bodelon, and Wyble (2002) have argued that theta oscillations help tune hippocampal dynamics for encoding vs. retrieval, such that dynamics are optimized for encoding during one phase of theta, and dynamics are optimized for retrieval during another phase of theta. The Hasselmo et al. (2002) model varies the relative strengths of different excitatory projections as a function of theta (to foster encoding vs. retrieval), but does not vary inhibition. Our impression is that our oscillating algorithm and Hasselmo's model are orthogonal rather than contradictory. As such, we may be able to combine the two models. One possibility would be to align the inhibitory oscillation (from our model) and the oscillation in excitatory projection strengths (from Hasselmo's model) such that inhibition is above-baseline during the 'encoding' phase of theta and inhibition is below-baseline during the 'retrieval' phase of theta. As per our theory, learning would be based on changes in activation triggered by changing inhibition. This

method of lining up the oscillations has the useful property that the oscillation phase primarily associated with memory strengthening in our model (high inhibition) matches up with the oscillation phase associated with LTP in the Hasselmo model ('encoding mode'), and the oscillation phase primarily associated with competitor punishment in our model (low inhibition) matches up with the oscillation phase associated with LTD in the Hasselmo model ('retrieval mode'). We will explore the viability of this combined model in future research.

3. Model of memory protection during REM

The preceding section focused on problems with the learning rules used by CLS models, and how these problems might be addressed using the oscillating learning algorithm. However, there are other, deeper issues with the CLS framework that cannot be addressed simply by changing the learning rule. In this section, we discuss the problem of *non-stationary environments*: How does the network maintain a representation of stimuli that temporarily drop out from the training set? We discuss how existing CLS models fail to solve this problem, and how this problem can be addressed by adding a new kind of off-line learning that rehearses and protects existing knowledge structures.[2]

3.1. Why we need two kinds of off-line learning

The original form of the Complementary Learning Systems framework as proposed by McClelland et al., 1995 included a single form of off-line learning in which hippocampus replayed memories to cortex. The role of this off-line learning was to allow the cortical model to incorporate information about rare events. As discussed below, this framework works well when the environment is stationary (i.e. the composition of the training set does not change) but it fails to preserve existing knowledge when the environment is not stationary. This point can be illustrated by considering what happens to our knowledge of typical birds after seeing a penguin. We will consider two situations: the *stationary environment* case (where the subject continues to see typical birds) and the *non-stationary environment* case (where typical birds are temporarily removed from the environment).

Penguin learning with a stationary environment. In this case, the person sees typical (winged, feathery, flying) birds on a regular basis during waking. One day, the person goes to the zoo and sees a (winged, feathery, flightless) penguin. The next day, the person returns to seeing typical birds. The original CLS model learns about penguins by taking a hippocampal 'snapshot' of the penguin, and then re-playing this memory to cortex. Hippocampal playback of 'penguins do not fly' will incrementally degrade the network's knowledge that (typically) birds fly. However, if the network continues to encounter typical birds (with high frequency) during waking, learning about these typical birds will repair the damage done by off-line learning about penguins.

Penguin learning with a non-stationary environment. In this case, the person sees typical birds on a regular basis during waking. Then, the person takes a month-long trip to Antarctica during which they only see penguins (never typical birds). In this case, hippocampal playback of new penguin memories and repeated environmental exposure to penguins will degrade the network's knowledge about typical birds. Because (in this example) typical birds are not present in Antarctica, learning-during-waking will not help repair the network's knowledge. The only possible source of support for typical birds in this situation is hippocampal replay of 'typical bird' memories from before the trip. However, as time passes, new information will be encountered in Antarctica that will also require off-line playback. Gradually, the probability that pre-trip information will be replayed, relative to Antarctica memories, will become extremely small. Ultimately, when neither the environment nor the hippocampus provides cortex with additional exposure to pre-trip information, it will fade from cortex.[3]

3.2. Lessons from the penguin example

The penguin example illustrates that (in the original form of CLS) the environment is responsible for repairing damage to existing knowledge. When existing knowledge continues to be reinforced by stimuli in the environment, CLS does fine. But, if the environment changes (such that existing knowledge is no longer directly supported by the environment) then the network will show high levels of interference.

The fact that the network shows some forgetting of typical birds, in and of itself, is not damning: From a computational perspective, it is appropriate to decrease the prominence of flying (vs. flightless) birds in semantic memory if the base rate of encountering flying birds decreases. However, the excessive speed of forgetting exhibited by the CLS model (and other models like it) is highly problematic. Taken literally, this property of the CLS model would imply that a person who regularly spends summers in Antarctica and the rest of the year in New Jersey would forget everything about New Jersey when they go to Antarctica, and vice-versa. To address the problem of catastrophic forgetting in non-stationary environments, we suggest that a second off-line learning mechanism is needed. The role of the second mechanism would be to slow the rate of erosion of pre-existing memories. This mechanism needs to be able to strengthen memories in situations where they are not being supported by the environment. In the next section, we discuss how learning during REM sleep may help to protect

[2] The simulation work described in this section was conducted as part of Adler Perotte's senior thesis research at Princeton University.

[3] The above argument is based on the idea that, as the person spends more and more time in Antarctica, the ratio of Antarctica episodic memories to pre-trip episodic memories will increase, resulting in proportionately less rehearsal of pre-trip episodic memories. This prediction depends critically on the rules that govern which memories get replayed by the hippocampus. It is possible (in principle) that one could devise a clever algorithm for hippocampal replay that continues to give privileged status to pre-trip memories. However, in practice, we are not sure how this goal could be accomplished.

memories, and we present a neural network model of this process.

3.3. Data on sleep and learning

The need for two distinct kinds of off-line learning (hippocampal replay of new memories to cortex, and repair of pre-existing memories, respectively) converges strongly with recently acquired data on sleep and learning (for reviews, see Gais & Born, 2004; Paller & Voss, 2004; Ribeiro & Nicolelis, 2004; Stickgold, 1998; Walker & Stickgold, 2004). These findings suggest that slow wave sleep (SWS) and REM sleep contribute to learning in distinct ways: SWS may support hippocampal replay of new memories to cortex, and REM may support tuning of pre-existing cortical and hippocampal representations. We briefly review the evidence for this linkage below.

Evidence linking SWS to hippocampal replay. The strongest evidence for hippocampal replay during SWS comes from electrophysiological studies that have examined the relationship between hippocampal activity in SWS vs. waking. Several studies have found that patterns of neural activity observed during waking events reappear in subsequent periods of sleep, and this replay occurs more frequently in SWS than in REM (see, e.g. Wilson & McNaughton, 1994 for evidence of replay in SWS; but see Louie and Wilson, 2001 for evidence that some replay occurs during REM; see Ribeiro et al., 2004 for a direct comparison showing more replay in SWS than REM). Other studies have found that, during SWS, hippocampal replay of memories is coherent with cortical reactivation (Qin, McNaughton, Skaggs, & Barnes, 1997). Also, sharp wave-ripple activity in hippocampus has been shown to be temporally correlated to sleep spindle oscillations in cortex during SWS (Siapas & Wilson, 1998; Sirota, Scicsvari, Huhl, & Buzsaki, 2003). Finally, Hasselmo (1999) observed that acetylcholine levels in the hippocampus are lower during SWS vs. waking and REM. Hasselmo (1999) goes on to describe how low acetylcholine levels should facilitate retrieval of stored hippocampal memories (e.g. by increasing the relative strength of CA3 recurrents). Although there is much work to be done in specifying the exact nature of the hippocampo–cortical interaction during SWS, these findings are broadly consistent with the idea that (during SWS) hippocampus is teaching cortex about recent events. For additional discussion of this point, see Buzsaki (1998); Gais and Born (2004); Hasselmo (1999), and Sejnowski and Destexhe (2000); for computational models of this process, see Alvarez and Squire (1994); Meeter and Murre (in press).

Evidence linking REM to neural plasticity. There is extensive evidence, both direct and indirect, suggesting that REM sleep plays an important role in neural plasticity. For example, theta oscillations, which have been correlated to human memory formation (e.g. Sederberg, Kahana, Howard, Donner, & Madsen, 2003), are prevalent during REM sleep (Winson, 1993). On a cellular level, Ribeiro and Nicolelis (2004) show that transcriptional factors, associated with the formation of memories during waking, are up-regulated during REM. Also, recent studies have found behavioral evidence that directly relates REM to learning on non-declarative memory tasks. For example, Smith, Nixon, and Nader (2004) found that the number and density of rapid eye movements (REMs) increased, relative to a control group, after subjects performed difficult novel tasks (mirror tracing and tower of Hanoi). Additionally, the number of REMs correlated with the degree of improved performance following sleep.

Importantly, while several studies have found evidence for hippocampo–cortical interactions during SWS, there is much less evidence for hippocampo–cortical synchrony during REM. For example, while theta oscillations are more prevalent in REM than SWS, these oscillations are not synchronized between hippocampus and cortex (Cantero, Atienza, Stickgold, Kahana, Madsen and Kocsis, 2003). These results suggest that REM involves separate learning processes occurring within cortex and hippocampus, as opposed to transfer of information from hippocampus to cortex (for additional discussion of how REM could tune cortical representations, see, e.g. Hasselmo, 1999).

3.4. The REM sleep model

In this section, we provide a brief overview of our model of REM sleep. Based on the data reviewed above, it seems safe to conclude that some kind of learning occurs during REM. However, it is not clear (based on this data) how REM achieves the functional goal of repairing damaged memories. The goal of the modeling work presented here is to bridge the gap, and show (to a first approximation) how a process with the physiological properties described above can support memory protection and repair.[4]

The most critical functional properties of REM, as reviewed above, are: (1) cortical neural activity is unaffected by environmental stimuli and uncorrelated with hippocampal activity, and (2) theta oscillations are prevalent. As such, we have modeled REM sleep as a period in which the cortical and hippocampal networks are dissociated from external input (and from each other) and autonomously rehearse stored memories. In keeping with the finding of strong theta activity during REM, learning during REM in our model is guided by inhibitory oscillations (as per the *Oscillating Learning Algorithm* section above).

The simulation of REM sleep presented here uses a cortical memory architecture. We discuss later how to re-integrate the hippocampal network into the model. With respect to cortical learning, we view REM as a period where cortex can 'think about what it already knows', thereby reinforcing knowledge that may no longer be supported by the environment or by the hippocampus. In our model, we initiate REM rehearsal by presenting the cortical network with a single noisy input and allowing the network to activate a memory. Once the REM

[4] Note that our model of the REM sleep process should not be confused with the Shiffrin and Steyvers (1997) REM model of recognition memory, which has nothing to do with REM sleep.

episode is initiated, no further input is given to the network. During REM rehearsal, the network transitions from one attractor state to another due to synaptic depression, which temporarily weakens the active pattern.

The model's ability to repair damaged memories depends critically on non-linear attractor dynamics in the cortical network. These attractor dynamics give the network the ability to recall the intact version of a memory even if the synapses underlying that memory have been disrupted. There are clearly limits to this dynamic: After a certain amount of damage, the memory will simply cease to exist as an attractor state in the cortical network. However, there is a relatively large window where disrupting the underlying synaptic substrate of a memory does not compromise recall of the pattern. This is analogous to a building where the support beams are crumbling but the building is still standing. If the REM rehearsal process succeeds in finding a memory in this state (i.e. where it can still be recalled, but the synaptic substrate is weak), then the memory can be repaired.

In our model, learning during REM uses the same oscillation-based learning algorithm that was used in the pattern completion and familiarity simulations presented earlier. As memories are rehearsed, the model oscillates the strength of inhibition, and changes weights based on changes in activation triggered by these inhibitory oscillations. We discussed earlier how raising inhibition allows us to 'stress-test' a memory. If a memory is weak (because of damage incurred during SWS or awake learning, or because of inadequate training), it will show decreased activity when we raise inhibition, which, in turn, will trigger learning processes that strengthen the memory. At an intuitive level, one can think of SWS and awake learning as 'denting' existing memories (like you would dent a car) but not destroying them; REM learning provides a way of repairing these dents. The second component of the learning algorithm (weakening memories that activate when inhibition is lowered) also plays a critical role. This competitor punishment mechanism allows existing memories to push away new memories that are encroaching on their space. More generally, this mechanism works to ensure that memories retain their individuating features and do not all collapse together (a problem that affects other models of memory consolidation; see the *Preventing runaway consolidation* section for additional discussion of this point). The idea that competitor punishment occurs during sleep also leads to specific behavioral predictions regarding effects of sleep on memory, as discussed in the *Applications to specific findings* section below.

3.5. Simulation: AB–AC interference

We used a simple list learning paradigm to explore how incorporating REM affects learning and forgetting in our cortical model. In particular, we were interested in exploring how REM affects learning in situations where the environment is not stationary. The network architecture consisted of three 50-unit layers (input, output, and hidden). For this set of simulations, the input layer was bidirectionally connected to the hidden layer, and the hidden layer was bidirectionally connected to the output layer; there were no recurrent connections within layers. Inhibition was oscillated on the input and output layers. The training patterns for our model were created in the spirit of the McCloskey and Cohen (1989) AB–AC model of catastrophic forgetting. The 'AB' list consisted of 15 randomly generated input–output pairs. The 'AC' list was generated by taking each pattern from the AB list and changing three units (out of 5) from both the input and output patterns such that each new pattern was 40% similar to the corresponding AB pattern. This high level of similarity between the two lists made it difficult for the network to maintain memories of the AB items as it learned the AC items.

The network was initially trained on the AB items to criterion. Next, we started training the network on AC patterns. For these trials, the AC pattern was presented directly to the network. These AC trials can be viewed as a proxy for learning occurring during waking and SWS. In the *REM Sleep* condition, the network was allowed to do REM rehearsal after each epoch of AC training. In the *No REM Sleep* condition, the network was not allowed to do any REM rehearsal in between AC epochs. After each epoch of AC training, we tested the network's memory for all of the AB and AC patterns by presenting the input-layer pattern and measuring the network's ability to recall the corresponding output-layer pattern (note that learning was turned off during test trials, and inhibition was not oscillated at test). By comparing the REM Sleep and No REM Sleep conditions, we were able to explore how including REM affects retention of AB items and learning of AC items.

3.6. Implementation of REM

Basic REM parameters. The learning parameters used during REM were identical to the learning parameters used during AB and AC learning, except for the fact that we used a smaller inhibitory oscillation size during REM vs. AB and AC learning. The most important difference between REM vs. AB and AC learning is that external patterns were not applied to the network during REM. Each REM episode was started by initializing the activity values of the network to a random value. After this initial burst of noise, the network was allowed to autonomously generate patterns to rehearse. Each REM episode lasted for 30,000 time steps. This value was selected because it allowed the network to visit and repair enough representations during REM to stabilize memories from the AB list. The oscillating learning algorithm was run continuously through the REM episode. Weight change values were computed on a time-step-by-time-step basis during the REM episode, but the weight changes were not actually implemented until the end of the REM episode.

Preventing runaway consolidation. An important problem that autonomous rehearsal mechanisms need to solve is *runaway consolidation* (Ans & Rousset, 2000; Meeter, 2003; Wittenberg, Sullivan, & Tsien, 2002). This problem arises when some memories are stronger than others. When random noise is injected into the system, the probability of recalling a

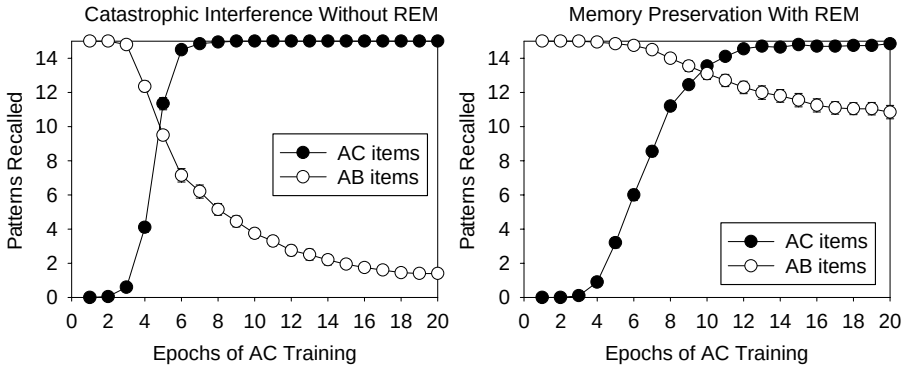

Fig. 4. Graphs of how AC training affects memory for AB and AC items, both with and without REM. Each graph plots the number of AB and AC items correctly recalled, after each epoch of AC training. The left-hand graph shows the model's performance without REM. The right-hand graph shows the model's performance with REM. A comparison of the two figures shows that including REM episodes greatly reduces the forgetting of the AB items, at the cost of slightly slowing acquisition of AC items.

memory is a function of that memory's strength. Thus, strong memories are rehearsed more often than weak memories. This leads to a positive feedback loop: Patterns that are rehearsed become even stronger, which makes them even more likely to be rehearsed in the future. This pattern of rehearsal leads to a situation where a small number of memories become extremely strong, and all other memories in the system become extremely weak. Rehearsal algorithms that manifest this problem are obviously unsuited for the task of preserving stored knowledge.[5]

To prevent the network from repeatedly settling into the same patterns, we implemented a synaptic depression mechanism. This mechanism slowly reduces the efficacy of connections between concurrently active units. As time progresses, the active pattern tires and dissipates. When this happens, other units activate and the network settles into a new pattern. Once a depressed connection is no longer being used, it begins to rebuild its strength. We selected a synaptic depression mechanism that depends on both presynaptic and postsynaptic activity, rather than a mechanism that depends only on presynaptic activity (e.g. Huber & O'Reilly, 2003; Gotts & Plaut, 2002), because the former mechanism is much more specific in targeting the active memory. When depression depends entirely on presynaptic activity, it will generalize to all memories that share neurons with the active memory, whereas depression that depends on presynaptic and postsynaptic activity will only generalize to (the smaller set of) memories that share synapses with the active memory. Having said this, however, the basic pattern of results reported in the next section does not depend on our use of the 'post+pre' synaptic depression mechanism; the same qualitative pattern was found when we used depression based on presynaptic activity only.

[5] To solve the problem of runaway consolidation, it is not necessary to completely eliminate effects of memory strength on rehearsal. We suspect that this is not possible, nor would it be desirable in light of behavioral data suggesting that (during awake learning of word lists) strong items are rehearsed more often than weak items (e.g. Ward, Woodward, Stevens, & Stinson, 2003). Rather, the goal is to ensure that weak memories continue to be rehearsed, to a degree that is sufficient to preserve these memories.

3.7. Learning with and without REM

The inclusion of the REM episodes after each epoch of AC training greatly reduced the rate of forgetting of the AB items. This can be seen in Fig. 4. Without REM, the average number of AB patterns recalled dropped below 2 (out of 15) after 20 epochs of AC training. In contrast, with the inclusion of REM, the network was able to retain more than 11 of the AB items after 20 epochs of AC training. In addition to reducing forgetting of AB items, REM also slowed down acquisition of AC items. The network was able to learn all 15 AC patterns both with REM and without REM, but this process took approximately 12 epochs of training with REM, vs. 6 epochs without REM. The slower pace of learning with REM reflects the fact that weaving new memories in with old memories (without destroying the old memories) is a more demanding process than simply letting the new memories overwrite the old memories. In the former case, the network has to shuffle around representations to make room for both AB and AC memories, whereas in the latter case the network can simply re-use the same set of neurons. Because REM model learned the same number of AC memories and retained more AB memories, the total number of patterns stored in the network at the end of training was larger with REM than without REM (26 vs. 16).

3.8. REM discussion

In summary: Adding 'REM sleep' periods to the McClelland et al. (1995) Complementary Learning Systems model significantly reduces the amount of forgetting. In this section, we briefly review how our model relates to other theoretical accounts of memory protection, and we discuss future directions for the model.

3.9. Relation to other computational models of memory protection

Our model of how REM preserves memories can be viewed as a descendant of models proposed by French (1997) and Ans and Rousset (1997), and later by Ans and Rousset (2000).

These studies pioneered the use of random noise to elicit (and then learn about) stored memory patterns (see also Wittenberg et al., 2002). The main difference between our model of cortical memory preservation and the Ans and Rousset (2000) model is that Ans and Rousset (2000) use two networks (with basically identical properties) to implement cortical memory preservation, whereas our model uses a single, unified network. The second network in their model maintains pristine copies of older memories, which the first network can subsequently use to repair representations that were damaged in new learning. The primary contribution of our work is to show that damage caused by new learning can be repaired without consulting an undamaged copy of the knowledge base. As discussed above, our scheme exploits the fact that—when synaptic weights have been disrupted by a relatively small amount—it is still possible to retrieve the memory in its original form. Thus, so long as the damaged memory is located (during REM rehearsal) before it becomes unrecallable, it can be repaired. This allows us to dispense with the neurobiologically implausible 'second cortical network' posited by Ans and Rousset (2000).

Other solutions to the stability–plasticity problem, such as Carpenter and Grossberg's Adaptive Resonance Theory (ART; Carpenter & Grossberg, 1988, 2002, 2003), do not require off-line learning. ART avoids catastrophic interference by gating when learning occurs. The gating mechanism prevents learning when the current pattern differs too much from top-down expectations (and, thus, learning the current pattern would significantly alter these top-down expectations). Although it is impressive that ART does not require off-line learning, this functional strength can also be viewed as an explanatory shortcoming: Because ART does not need off-line learning, it does not provide a natural explanation for data showing that off-line learning (during sleep) actually occurs.

One other model of note is the cortico–hippocampal model published by Kali and Dayan (2004). In this paper, Kali and Dayan (2004) present simulations showing that hippocampal replay alone (in the absence of an extra 'memory protection' process) can help to preserve semantic memories after the statistics of the training environment change. On the surface, this result appears to be inconsistent with our claim that an extra memory protection process is required to fully address the catastrophic interference problem. However, the utility of the Kali and Dayan (2004) model (with regard to solving the catastrophic interference problem) is compromised by two issues. First, the simulations presented in Kali and Dayan (2004) only explore the effect of subsequent cortical learning on memory storage, not the effect of subsequent hippocampal learning (specifically: no new hippocampal memories are formed after the statistics of the training environment change). Furthermore, the hippocampal component of the model is not explicitly simulated, so they cannot explore the possibility that new hippocampal memories might distort previously stored hippocampal memories. We suspect that updating the Kali and Dayan (2004) model to address these issues (by adding new episodic learning after the training environment changes, and allowing for the possibility of interference within the hippocampus) would greatly compromise their model's ability to preserve stored semantic knowledge.

3.10. Future directions

In this section, we have provided a simple demonstration of how adding a 'REM sleep' mechanism to CLS can help to minimize interference. That said, there is a great deal of work that remains to be done in understanding the neural mechanisms that support off-line learning (and their functional consequences). In this section, we describe ways in which the model can be refined and extended, and ways in which the model can be applied to specific sleep and learning findings.

Re-integrating the hippocampal model. Having demonstrated the basic properties and feasibility of the REM memory protection mechanism (as applied to cortex), the next logical step is to add the hippocampal network from Norman and O'Reilly (2003) back into the model. Re-integrating the hippocampal model would allow us to explicitly model waking, SWS, and REM sleep. During waking, the hippocampus would learn with a very high learning rate, and cortex would learn with a much smaller learning rate (as per basic CLS principles). During SWS, the hippocampal network would recall memories acquired during the waking state through a random settling process (similar to that used during REM in the cortical model), and cortex would learn based on these hippocampal training trials. As per the ideas described in Hasselmo (1999), we would adjust modulatory parameters in the hippocampal model to facilitate retrieval during SWS (vs. encoding during waking and REM). The model would be configured to stay in SWS long enough to sample recently acquired hippocampal memories, but not so long that SWS destroys the attractor network of the cortical model. After SWS, we would implement the REM memory protection process in both the cortex and the hippocampus independently. Although we used a cortical architecture in the simulations described above, the same basic principles of autonomous re-activation and strengthening can also be applied to the hippocampal model. As mentioned earlier, although there is less overlap between memories in hippocampus vs. cortex, there is still *some* overlap, which leads to interference. If enough interference builds up, this could prevent the hippocampus from fulfilling its job of conveying recent memories to cortex during SWS. As such, the hippocampus (like cortex) stands to benefit from the memory protection mechanisms discussed in this section.

Targeting damaged memories. In the REM simulation presented in this paper, the REM rehearsal process was able to sample (almost) all of the AB traces because there were only 15 of these traces. In a brain-size network, with thousands (or millions) of attractor states, this kind of exhaustive sampling is not possible. In this situation, the REM rehearsal process needs to be able to selectively sample the relatively small set of memories that have suffered the most damage during SWS. A major future direction for the REM model is to explore mechanisms that will promote selective sampling of damaged (vs. non-damaged) memories.

There are a number of potential solutions to this problem. One possibility is to incorporate a weak influence of the hippocampus during the process of REM. According to this view, the hippocampal network would continue to recall memories, but (unlike SWS) this hippocampal influence would not be strong enough to force a pattern of activity on the cortical network. Rather, its influence on cortex would serve to weakly guide activity to the areas of attractor space that were visited during SWS (and, therefore, were most likely to have been damaged). The weak hippocampal input would trigger re-activation of cortical attractors in the 'damaged' areas, thereby making it possible to repair these attractors. Another, related possibility (which does not require hippocampo–cortical interactions during REM) would be to apply a hysteresis algorithm to units visited during SWS. This hysteresis algorithm would enact a temporary increase in the efficacy of neurons and/or synapses that were activated during SWS. If hysteresis carried over (from SWS) into the following REM phase, it would serve to guide the cortical network to regions of attractor space that were visited during SWS.

One possible issue with both the 'weak hippocampal influence' and 'hysteresis' ideas is that, by guiding cortex to regions of attractor space visited during SWS, these mechanisms may foster additional strengthening of new memory traces, as opposed to repair of old memory traces. However, we do not think this is a major concern, for two reasons: First, if a memory is truly new, then its cortical memory trace will be weaker than the cortical memory traces of pre-existing memories, so cortex will be more likely to rehearse the pre-existing memories. Second, even if the network does rehearse new memories (to some extent) during REM, eventually these memories will tire out due to synaptic depression. If the network stays focused on the same region of attractor space (and new memories are depressed) then it will rehearse old memories in that region.

Applications to specific findings. Another future direction is to use the model to simulate specific sleep-and-learning datasets. Over the past few years, several studies have been published that go beyond proving the mere existence of learning during sleep, and map out a more detailed landscape of how sleep affects learning. In this section, we will sketch out how our ideas about memory competition (and competitor punishment) during REM can be applied to some puzzling data from Walker, Brakefield, Hobson, and Stickgold (2003) on how sleep affects memory for motor sequences.

The basic finding from this study is that sleep enhances memory for simple motor sequences in a button-pressing task. Walker et al. (2003) build on this finding in several different ways. In one variant of this paradigm, subjects learned one sequence (S1) and then learned a second sequence (S2) immediately afterward. Memory for S2 (measured in terms of accuracy) improved after sleep, but memory for S1 did not. However, when six (waking) hours intervened between learning S1 and S2, both sequences showed improved accuracy after sleep. Walker et al. (2003) explain this finding in terms of the idea that 6 h of waking can 'stabilize' a memory, thereby protecting it from interference from subsequent learning. However, this pattern of results can also be explained in terms of competitive dynamics during REM. In the 'no delay' condition, the two memories are encoded in a very similar spatiotemporal context. This contextual overlap makes the memory traces associated with S1 and S2 more similar, which in turn increases the extent to which the two memories compete during sleep. Since S2 is stronger, it is more likely to win the competition (and S1 is more likely to lose), which implies that S2 will benefit more from REM than S1. In contrast, when 6 h intervene between learning the sequences, the two sequences will be associated with relatively distinct sets of contextual features (e.g. you might be hungry when learning S1 but not when learning S2). As a result, the cortical engrams of S1 and S2 will be more different in this condition than in the 'no delay' condition. Because there is less overlap, the memory traces are less likely to compete, so they both should benefit equally from REM.

In a related finding, Walker et al. (2003) trained subjects on S1 and let them sleep (so S1 performance improved). On the second day of the experiment, Walker et al. (2003) trained subjects on S2. Prior to learning S2, some subjects were briefly re-exposed to S1, and some subjects were not. On the third day, all subjects showed improved memory for the S2. However, the re-exposure manipulation had a large effect on memory for S1: Subjects who were re-exposed to S1 showed a large decrease in S1 performance from day 2 to day 3. In contrast, subjects who were not re-exposed to S1 did not show a change in S1 performance. Walker et al. (2003) interpret these results in terms of *reconsolidation*; according to this idea, reactivating a memory temporarily makes the molecular substrate of that memory more labile, and thus more vulnerable to interference. For example, in the animal literature, several studies have found that 'reminding' an animal of a tone-shock association (by presenting the tone by itself) makes that tone-shock memory vulnerable to disruption via injection of a protein synthesis inhibitor (e.g. Nader, Schafe, & LeDoux, 2000; for a recent review, see Dudai & Eisenberg, 2004; for additional discussion of how the Walker et al., 2003 finding relates to the animal reconsolidation literature, see Nader, 2003). However, we can explain these results without positing re-labialization of the S1 memory. Rather, according to our REM framework, presenting S1 in the same context as S2 makes it more likely that S1 will pop up as a competitor to S2 during the second night of sleep, which (in turn) will hurt memory for S1.

We still need to build a working simulation of the Walker et al. (2003) data, in order to test the sufficiency of these ideas. However, even without a working simulation, it is not difficult to generate testable predictions that follow from our competitor-punishment account of these results. According to our theory, the key variable that determines whether S1 memory improves during sleep is the similarity of S1 to S2. This implies that, if we could make the S1 memory trace more different from the S2 trace (by changing the stimuli or motor responses, or by having subjects learn the two sequences in different rooms), this would allow the network to rehearse both S1 and S2 without them interfering with one another. In contrast, the reconsolidation account does not intrinsically

make predictions about effects of S1–S2 similarity (although it is not incompatible with the idea that higher similarity will lead to higher interference).

The fact that we can account for the Walker et al. (2003) data in terms of competitor punishment has led us to consider whether we can account for other reconsolidation findings in terms of competitor punishment. Specifically, in the fear conditioning paradigm described above (where subjects are exposed to a tone-shock pair, and later are 'reminded' of this association by presenting the tone), it is possible to construe the 'tone alone' presentation as a competitor to the original memory (i.e. tone is now being paired with safety, not shock), rather than a reminder of the original memory (see Eisenberg, Kobilo, Berman, & Dudai, 2003 for a similar idea). Given this premise, we can construct a competitor-punishment account of basic reconsolidation findings:

- One of the key ideas presented in this paper is that catastrophic interference is the 'default mode' for neural networks: In the absence of special memory protection mechanisms (e.g. the REM sleep mechanism presented here), new learning will weaken similar pre-existing memories. Rehearsal of the new memory during SWS will compound this effect, resulting in worse and worse recall of the pre-existing memory over time.
- In the fear conditioning paradigm, after the animal is exposed to the 'tone-safety' association, we posit that REM learning (or something like it) is needed to protect the original 'tone-shock' memory.
- Protein synthesis blockers may disrupt REM learning (see Ribeiro & Nicolelis, 2004 for discussion of transcriptional factors in REM sleep) while leaving other forms of neural plasticity relatively intact. In this situation (where learning about the tone-safety event is still taking place, in the absence of REM memory protection), the tone-safety memory will catastrophically interfere with the tone-shock memory, resulting in diminished fear conditioning.

At this point, these ideas are highly speculative (especially the idea that protein synthesis blockers would selectively disrupt REM memory protection mechanisms). However, in light of recent attempts to explain human learning data in terms of reconsolidation theory, we think it is equally important to consider whether animal 'reconsolidation' findings can be explained by network-level interference theories such as CLS. At the very least, this leads one to consider factors that were previously neglected (e.g. the exact relationship between the 'reminder' and the original memory; and also the role of different sleep stages in promoting/preventing reconsolidation effects).

3.11. Final thoughts

In recent years, Complementary Learning Systems research has focused on simulating specific findings (e.g. Norman & O'Reilly, 2003; O'Reilly & Rudy, 2001). While this approach has been very productive, it is also important to take a step back and assess how well CLS does at addressing the problem that it was designed to solve: accurately storing and maintaining knowledge about the environment. In this paper, we have outlined the challenges that the CLS model faces, in order to provide a satisfactory solution to the stability–plasticity problem. We have also tried to outline some possible ways of addressing these challenges. We showed how leveraging inhibitory oscillations can help reduce interference during learning, by ensuring that synapses are modified judiciously (i.e. such that strengthening is focused on weak target units, and weakening is focused on strong non-target units). We also showed how a 'REM sleep' process can be used to protect memories when these memories are no longer being directly supported by the environment. Importantly, although our discussion of the oscillating learning algorithm and REM sleep has focused on the cortical model, we think that the oscillating algorithm and the REM memory protection mechanism may also be applicable to the hippocampus. We are presently exploring this possibility.

The analyses presented here illustrate the vast distance that CLS has to travel before it solves stability–plasticity (e.g. we need to devise a mechanism that will allow the REM sleep model to scale up to larger networks, with more stored memories). However, CLS has a long history of turning its weaknesses into strengths: Understanding that networks with distributed, overlapping representations perform poorly at rapid memorization led to a better understanding of why we need a hippocampus, and how it works. Understanding that the standard CLS model fails to deal properly with non-stationary environments may lead to a better understanding of REM sleep. Finally, understanding the limitations of simple Hebbian learning may lead to a better understanding of the functional role of theta oscillations. This history gives us hope that—as we continue to chip away at stability–plasticity—our efforts will be repaid with further insights into the functional and neural architecture of learning.

References

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive Science*, *9*, 147–169.

Aggleton, J. P., & Brown, M. W. (1999). Episodic memory, amnesia, and the hippocampal-anterior thalamic axis. *Behavioral and Brain Sciences*, *22*, 425–490.

Alvarez, P., & Squire, L. R. (1994). Memory consolidation and the medial temporal lobe: A simple network model. *Proceedings of the National Academy of Sciences, USA*, *91*, 7041–7045.

Anderson, M. C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language*, *49*, 415–445.

Anderson, M. C., & Bell, T. (2001). Forgetting our facts: the role of inhibitory processes in the loss of propositional knowledge. *Journal of Experimental Psychology: General*, *130*(3), 544–570.

Ans, B., & Rousset, S. (1997). Avoiding catastrophic forgetting by coupling two reverberating neural networks. *Academie des Sciences, Sciences de la vie*, *320*, 989–997.

Ans, B., & Rousset, S. (2000). Neural networks with a self-refreshing memory: knowledge transfer in sequential learning tasks without catastrophic forgetting. *Connection Science*, *12*(1), 1–19.

Blaxton, T. A., & Neely, J. H. (1983). Inhibition from semantically related primes: Evidence of a category-specific retrieval inhibition. *Memory and Cognition*, *11*, 500–510.

Bogacz, R., & Brown, M. W. (2003). Comparison of computational models of familiarity discrimination in the perirhinal cortex. *Hippocampus*, *13*, 494–524.

Burgess, N., & O'Keefe, J. (1996). Neuronal computations underlying the firing of place cells and their role in navigation. *Hippocampus*, *6*, 749–762.

Buzsaki, G. (1998). Memory consolidation during sleep: A neurophysiological perspective. *Journal of Sleep Research*, *7*, 17–23.

Buzsaki, G. (2002). Theta oscillations in the hippocampus. *Neuron*, *33*, 325–340.

Cantero, J. L., Atienza, M., Stickgold, R., Kahana, M. J., Madsen, J. R., & Kocsis, B. (2003). Sleep-dependent theta oscillations in the human hippocampus and neocortex. *Journal of Neuroscience*, *23*, 10897–10903.

Carpenter, G. A., & Grossberg, S. (1988). The art of adaptive pattern recognition by a self-organizing neural network. *Computer*, *21*(3), 77–88.

Carpenter, G. A., & Grossberg, S. (2002). A self-organizing neural network for supervised learning, recognition, and prediction. In T. A. Polk, & C. M. Seifert (Eds.), *Cognitive modeling* (pp. 289–314). Cambridge, MA: MIT Press.

Carpenter, G. A., & Grossberg, S. (2003). Adaptive resonance theory. *The handbook of brain theory and neural networks* (pp. 87–90). Cambridge, MA: MIT Press.

Ciranni, M. A., & Shimamura, A. P. (1999). Retrieval-induced forgetting in episodic memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 1403.

Dudai, Y., & Eisenberg, M. (2004). Rites of passage of the engram: Reconsolidation and the lingering consolidation hypothesis. *Neuron*, *44*, 93–100.

Eichenbaum, H., Otto, T., & Cohen, N. J. (1994). Two functional components of the hippocampal memory system. *Behavioral and Brain Sciences*, *17*(3), 449–518.

Eisenberg, M., Kobilo, T., Berman, D. E., & Dudai, Y. (2003). Stability of retrieved memory: Inverse correlation with trace dominance. *Science*, *301*, 1102–1104.

French, R. M. (1997). Pseudo-recurrent connectionist networks: An approach to the 'sensitivity-stability' dilemma. *Connection Science*, *9*, 353–379.

French, R. M. (1999). Catastrophic forgetting in connectionist networks: Causes, consequences and solutions. *Trends in Cognitive Sciences*, *3*(4), 128–135.

French, R. M. (2003). Castrophic forgetting in connectionist networks. In L. Nadel (Ed.), *Encyclopedia of cognitive science*. London: MacMillan.

Gais, S., & Born, J. (2004). Declarative memory consolidation: Mechanisms acting during human sleep. *Learning and Memory*, *11*, 679–685.

Gotts, S. J., & Plaut, D. C. (2002). The impact of synaptic depression following brain damage: A connectionist account of 'access/refractory' and 'degraded-store' semantic impairments. *Cognitive, Affective, and Behavioral Neuroscience*, *2*(3), 187–213.

Grossberg, S. (1976). Adaptive pattern classification and universal recoding. I. Parallel development and coding of neural feature detectors. *Biological Cybernetics*, *23*, 121–134.

Hasselmo, M. E. (1999). Neuromodulation: Acetylcholine and memory consolidation. *Trends in Cognitive Sciences*, *3*(9), 351–359.

Hasselmo, M. E., Bodelon, C., & Wyble, B. P. (2002). A proposed function for hippocampal theta rhythm: Separate phases of encoding and retrieval enhance reversal of prior learning. *Neural Computation*, *14*, 793–818.

Hasselmo, M. E., & Wyble, B. (1997). Free recall and recognition in a network model of the hippocampus: Simulating effects of scopolamine on human memory function. *Behavioural Brain Research*, *89*, 1–34.

Hinton, G. E. (1989). Deterministic Boltzmann learning performs steepest descent in weight-space. *Neural Computation*, *1*, 143–150.

Hinton, G. E., & McClelland, J. L. (1988). Learning representations by recirculation. In D. Z. Anderson (Ed.), *Neural Information Processing Systems, 1987*, 358–366.

Hinton, G. E., & Sejnowski, T. J. (1986). Learning and relearning in Boltzmann machines. In D. E. Rumelhart, J. L. McClelland, & PDP Research Group, *Foundations*. *Parallel distributed processing* (Vol. 1) (pp. 282–317). Cambridge, MA: MIT Press.

Holscher, C., Anwyl, R., & Rowan, M. J. (1997). Stimulation on the positive phase of hippocampal theta rhythm induces long-term potentiation that can be depotentiated by stimulation on the negative phase in area CA1 in vivo. *Journal of Neuroscience*, *17*, 6470.

Huber, D. E., & O'Reilly, R. C. (2003). Persistence and accommodation in short-term priming and other perceptual paradigms: temporal segregation through synaptic depression. *Cognitive Science*, *27*, 403–430.

Huerta, P. T., & Lisman, J. E. (1996). Synaptic plasticity during the cholinergic theta-frequency oscillation in vitro. *Hippocampus*, *49*, 58–61.

Hyman, J. M., Wyble, B. P., Goyal, V., Rossi, C. A., & Hasselmo, M. E. (2003). Stimulation in hippocampal region CA1 in behaving rats yields long-term potentiation when delivered to the peak of theta and long-term depression when delivered to the trough. *Journal of Neuroscience*, *23*, 11725–11731.

Kahana, M. J. (2001). Theta returns. *Current Opinion in Neurobiology*, *11*, 739–744.

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

Koutstaal, W., Schacter, D. L., & Jackson, E. M. (1999). Perceptually based false recognition of novel objects in amnesia: Effects of category size and similarity to category prototypes. *Cognitive Neuropsychology*, *16*, 317.

Louie, K., & Wilson, M. (2001). Temporally structured replay of awake hippocampal ensemble activity during rapid eye movement sleep. *Neuron*, *29*, 145–156.

Marr, D. (1971). Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society (London) B*, *262*, 23–81.

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.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks. *The sequential learning problem*. In G. H. Bowes, *The psychology of learning and motivation* (Vol. 24) (pp. 109–164). San Diego, CA: Academic Press.

McNaughton, B. L., & Morris, R. G. M. (1987). Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends in Neurosciences*, *10*(10), 408–415.

Meeter, M. (2003). Control of consolidation in neural networks; avoiding runaway effects. *Connection Science*, *15*(1), 45–61.

Meeter, M., & Murre, J. (in press). Tracelink: A model of amnesia and consolidation. *Cognitive Neuropsychology*.

Minai, A. A., & Levy, W. B. (1994). Setting the activity level in sparse random networks. *Neural Computation*, *6*, 85–99.

Moll, M., & Miikkulainen, R. (1997). Convergence-zone episodic memory: Analysis and simulations. *Neural Networks*, *10*, 1017–1036.

Movellan, J.R. (1990). Contrastive Hebbian learning in the continuous Hopfield model. In D. S. Tourtezky, G. E. Hinton, & T. J. Sejnowski (Eds.), *Proceedings of the 1989 connectionist models summer school* (pp. 10-17).

Nader, K. (2003). Re-recording human memories. *Nature*, *425*, 571–572.

Nader, K., Schafe, G. E., & LeDoux, J. E. (2000). Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. *Nature*, *406*, 722–726.

Norman, K.A., Newman, E.L., & Detre, G.J. (2004). Further predictions from a neural network model of retrieval-induced forgetting. *45th Annual Meeting of the Psychonomic Society. Minneapolis, MN*.

Norman, K. A., Newman, E. L., Detre, G. J., & Polyn, S. M. (2005). How inhibitory oscillations can train neural networks and punish competitors. (Technical Report 05-1). Princeton, NJ: Princeton University, Center for the Study of Brain, Mind, and Behavior.

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: A complementary-learning-systems approach. *Psychological Review*, *4*, 611–646.

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

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., & Norman, K. A. (2002). Hippocampal and neocortical contributions to memory: Advances in the complementary learning systems framework. *Trends in Cognitive Sciences*, *12*, 505–510.

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.

Paller, K. A., & Voss, J. K. (2004). Memory reactivation and consolidation during sleep. *Learning and Memory*, *11*, 664–670.

Qin, Y. L., McNaughton, B. L., Skaggs, W. E., & Barnes, C. A. (1997). Memory reprocessing in corticocortical and hippocampal neuronal ensembles. *Philosophical Transactions: Biological Sciences*, *352*, 1525–1533.

Raghavachari, S., Kahana, M. J., Rizzuto, D. S., Caplan, J. B., Kirschen, M. P., Bourgeois, B., et al. (2001). Gating of human theta oscillations by a working memory task. *Journal of Neuroscience*, *9*, 3175–3183.

Ribeiro, S., Gervasoni, D., Soares, E. S., Zhou, Y., Lin, S. C., Pantoja, J., et al. (2004). Long-lasting novelty-induced neuronal reverberation during slow-wave sleep in multiple forebrain areas. *PLoS Biology*, *2*, 126–137.

Ribeiro, S., & Nicolelis, M. A. L. (2004). Reverberation, storage, and postsynaptic propagation of memories during sleep. *Learning and Memory*, *11*, 686–696.

Rolls, E. T. (1989). Functions of neuronal networks in the hippocampus and neocortex in memory. In J. H. Byrne, & W. O. Berry (Eds.), *Neural models of plasticity: Experimental and theoretical approaches* (pp. 240–265). San Diego, CA: Academic Press.

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery, and Psychiatry*, *20*, 11–21.

Sederberg, P., Kahana, M. J., Howard, M. W., Donner, E. J., & Madsen, J. R. (2003). Theta and gamma oscillations during encoding predict subsequent recall. *Journal of Neuroscience*, *23*, 10809–10814.

Sejnowski, T. J., & Destexhe, A. (2000). Why do we sleep? *Brain Research*, *886*, 208–223.

Sherry, D. F., & Schacter, D. L. (1987). The evolution of multiple memory systems. *Psychological Review*, *94*(4), 439–454.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM — retrieving effectively from memory. *Psychonomic Bulletin and Review*, *4*, 145–166.

Siapas, A. G., & &Wilson, M. A. (1998). Coordinated interactions between hippocampal ripples and cortical spindles during slow-wave sleep. *Neuron*, *21*, 1123–1128.

Sirota, A., Scicsvari, J., Huhl, D., & Buzsaki, G. (2003). Communication between neocortex and hippocampus during sleep in rodents. *Neuroscience*, *100*(4), 2065–2069.

Smith, C. T., Nixon, M. R., & Nader, R. S. (2004). Posttraining increases in rem sleep intensity implicate rem sleep in memory processing and provide a biological marker of learning potential. *Learning and Memory*, *11*, 714–719.

Squire, L. R. (1992). Memory and the hippocampus: A synthesis from findings with rats, monkeys, and humans. *Psychological Review*, *99*, 195–231.

Standing, L. (1973). Learning 10,000 pictures. *Quarterly Journal of Experimental Psychology*, *25*, 207–222.

Stickgold, R. (1998). Sleep: Off-line memory reprocessing. *Trends in Cognitive Sciences*, *2*, 484–492.

Sutherland, R. J., & Rudy, J. W. (1989). Configural association theory: The role of the hippocampal formation in learning, memory, and amnesia. *Psychobiology*, *17*(2), 129–144.

Teyler, T. J., & Discenna, P. (1986). The hippocampal memory indexing theory. *Behavioral Neuroscience*, *100*, 147–154.

Toth, K., Freund, T. F., & Miles, R. (1997). Disinhibition of rat hippocampal pyramidal cells by GABAergic afferents from the septum. *Journal of Physiology*, *500*, 463–474.

Treves, A., & Rolls, E. T. (1994). A computational analysis of the role of the hippocampus in memory. *Hippocampus*, *4*, 374–392.

Walker, M. P., Brakefield, T., Hobson, J. A., & Stickgold, R. (2003). Dissociable stages of human memory consolidation and reconsolidation. *Nature*, *425*, 616–620.

Walker, M. P., & Stickgold, R. (2004). Sleep-dependent learning and memory consolidation. *Neuron*, *44*, 121–133.

Ward, G., Woodward, G., Stevens, A., & Stinson, C. (2003). Using overt rehearsals to explain the word frequency effects in free recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 196–210.

Wilson, M. A., & McNaughton, B. L. (1994). Reactivation of hippocampal ensemble memories during sleep. *Science*, *265*, 676–678.

Winson, J. (1993). The biology and function of rapid eye movement sleep. *Current Opinion in Neurobiology*, *3*, 243–248.

Wittenberg, G. M., Sullivan, M. R., & Tsien, J. Z. (2002). Synaptic reentry reinforcement based network model for long-term memory consolidation. *Hippocampus*, *12*, 637–647.

Wu, X., Baxter, R. A., & Levy, W. B. (1996). Context codes and the effect of noisy learning on a simplified hippocampal CA3 model. *Biological Cybernetics*, *74*, 159–165.

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language*, *46*, 441–517.