

Integrating incremental learning and episodic memory models of the hippocampal region

M. Meeter¹, C.E. Myers² & M.A. Gluck³

¹ Department of Cognitive Psychology, Vrije Universiteit Amsterdam

² Department of Psychology, Rutgers University

³ CMBN, Rutgers University

Key Words: memory, classical conditioning, hippocampus, medial septum, cerebellum

Correspondence should be addressed to:

M. Meeter
Dept. of Cognitive Psychology
Vrije Universiteit Amsterdam
Vd Boechorststraat 1
1081 BT Amsterdam
The Netherlands
m@meeter.nl / tel. *31-20-4448993.

Abstract

By integrating previous computational models of cortico-hippocampal function, we develop and test a unified theory of the neural substrates of familiarity, recollection, and classical conditioning. Our approach integrates models from two traditions of hippocampal modeling, those of episodic memory (e.g, Norman & O'Reilly, 2003) and incremental learning (e.g., Gluck & Myers, 1993), by drawing on an earlier mathematical model of conditioning, SOP (Wagner, 1981). Our model describes how a familiarity signal may arise from parahippocampal cortices, giving a novel explanation for the finding that the neural response to a stimulus in these regions decreases with increasing stimulus familiarity. Recollection is ascribed to the hippocampus proper. It is shown how the properties of episodic representations in the neocortex, parahippocampal gyrus and hippocampus proper may explain phenomena in classical conditioning. The model reproduces the effects of hippocampal, septal, and broad hippocampal region lesions on contextual modulation of classical conditioning, blocking, learned irrelevance, and latent inhibition.

Integrating incremental learning and episodic memory models of the hippocampal region

Many recent theories of hippocampal functioning are variants of the idea that the hippocampus, or more broadly the hippocampal region (hippocampus proper, dentate gyrus, subiculum, and entorhinal, perirhinal and parahippocampal cortex), stores episodic memories (Eichenbaum, 1992; Hasselmo & Wyble, 1997; Marr, 1971; McClelland & Goddard, 1996; Meeter & Murre, in press; Meeter, Talamini, & Murre, 2004; Norman & O'Reilly, 2003; Talamini, Meeter, Murre, Elvevåg, & Goldberg, in press). In its strictest definitions, 'episodic' refers to the memories for unstructured array of things that were present or events that occurred at one specific location and moment in time. Evidence that the hippocampus plays a role in storing such memories comes from electrophysiological recordings (Eichenbaum, 2000; Ferbintineau & Shapiro, 2003; O'Keefe, 1979), functional imaging (Cabeza & Nyberg, 2000), patient data (Reed & Squire, 1998; Rempel-Clower, Zola, Squire, & Amaral, 1996; Scoville & Milner, 1957), and lesion studies with experimental animals (Gilbert, Kesner, & Lee, 2001; Jarrard, 1995).

However, lesion data has also implicated the hippocampal region in tasks that do not conform very well to the episodic label. These findings, discussed below, suggest that the hippocampus has a role in what is usually referred to as associative learning (Pearce & Bouton, 2001) or incremental learning (Gluck, Meeter, & Myers, 2003). Both names are apt, as this domain involves tasks in which the learner is required to associate a set of stimuli or a state of the environment with a behavioral output. Moreover, the output usually develops over tens or hundreds of trials. This is a far cry from episodic memory tasks, which often require retrieval after just one presentation of a certain material.

Findings from the incremental learning literature have led to a second strain of theories of the hippocampal region. The hippocampal region is theorized to aid the slow development of the correct representations to support behavioral performance (Gluck & Myers, 1993; Schmajuk & DiCarlo, 1992). In classical conditioning, for example, this may take the form of developing representations that distinguish conditions predictive of an unconditioned stimulus, and conditions that predict no such stimulus will occur. The hippocampal region may thus slowly alter and adapt representations in response to environmental feedback (i.e., rewards, punishments, unconditioned stimuli).

The two groups of theories, those concerned with episodic memory and those addressing incremental learning, seem diametrically opposed in the demands they pose to hippocampal functioning. Incremental learning is sensitive to behavioral outcome, while episodic learning is thought of as unsupervised, automatic coding of whatever is present. Episodic learning is fast (often one trial), while incremental learning is slow. It is not immediately obvious how a model could reconcile episodic and incremental learning in a parsimonious way. Here, we present a model which attempts to do just that, reconcile episodic memory with incremental learning. The conceptual glue that binds these two approaches is an earlier mathematical model of conditioning, the Sometimes Opponent Processing or SOP model (Wagner, 1981). The model is able to account for the basic phenomena of episodic memory, as well as for classic findings in the conditioning literature. Moreover, it offers a novel explanation for a puzzling finding from neurophysiology, namely that familiarity decreases (rather than increases) neural responses to familiar stimuli (Li, Miller, & Desimone, 1993; Xiang & Brown, 1998).

INCREMENTAL LEARNING & EPISODIC MEMORY

In the remainder of this paper, we will first outline the two strains of models, and some of the evidence for the roles of the hippocampal region in different kinds of memory. Then we will present our model, both in its theoretical basis and applications to concrete data. Although it is applicable to a wide range of memory tasks, we will in this paper focus on what may be seen as two extremes: explicit episodic memory, and classical conditioning. The paper ends with a discussion of the merits and limitations of the model, a comparison with other models, and a list of untested predictions.

Role of hippocampal region in episodic memory

The role of the hippocampus in memory has been the focus of many qualitative theories and computational models (Eichenbaum, 1992; Hasselmo & Wyble, 1997; Marr, 1971; McClelland & Goddard, 1996; Meeter, Murre, & Talamini, 2002; Meeter et al., 2004; Norman & O'Reilly, 2003; Talamini et al., in press). These models have assumed that the hippocampal region simply stores whatever pattern is presented to it by the neocortex. Input from the neocortex is modeled as arbitrary vectors to be stored. The hippocampal region forms a compact code that is bidirectionally linked to such neocortical representations. If a partial cue can later reactivate this compact code, it can retrieve the neocortical representation in its entirety.

The hippocampus is not the only structure involved in episodic memory. Parahippocampal cortices adjacent to the hippocampus have also been implicated in memory, especially in the processing of stimulus familiarity (Aggleton & Brown, 1999; Davachi, Mitchell, & Wagner, 2003; Ranganath et al., 2004; Zhu, McCabe, Aggleton, & Brown, 1997). Areas of the neocortex, down to unimodal primary cortices, play a role too: retrieval of visual memories will not only activate the medial temporal lobe, but also visual areas down to primary visual area V1 (Cabeza & Nyberg, 2000; Reber & Squire, 1998). In line with these findings, lesions to visual neocortical areas in the occipital lobe can cause amnesia for visual details (Ogden, 1993).

Damage to the basal forebrain can also cause a dense amnesia, suggesting a role in memory (Tranel, Damasio, & Damasio, 2000). Within this set of structures, the importance of the medial septum for hippocampal functioning has often been stressed (Hasselmo & Bower, 1993; Hasselmo & Wyble, 1997; Meeter et al., 2004; Myers et al., 1996; Rokers, Mercado, Allen, Myers, & Gluck, 2002), as it is the main source of acetylcholine in the hippocampus (Alonso, Sang U, & Amaral). Because acetylcholine facilitates learning (Hasselmo, 1999), the medial septum may thus control the learning process within the hippocampus.

From these findings rises a view of memories as a set of rich, modality-specific representations stored in the neocortex, bound together by high-level compound representations stored in the hippocampus (Murre, 1996; Teyler & DiScenna, 1986). This distinction between representations of different complexity need not be viewed as a dichotomy; in fact, representations of intermediate complexity may exist in areas such as the parahippocampal gyrus. A natural view of episodic memories would thus be a hierarchy of representations all bound together, with low-level features stored in the neocortex and integrated by higher-level representations in the hippocampal region (Talamini et al., in press). The septum may fit in this story as the controller of the learning rate in the hippocampus.

Role of the hippocampal region in incremental learning

In classical conditioning, subjects learn that a previously neutral stimulus (the conditioned stimulus or CS) precedes a response-evoking stimulus (the unconditioned stimulus or US). With time, they learn to give an anticipatory or preparatory response (the conditioned response or CR) to the CS. Such conditioning can be analyzed rather successfully as the learning of relations between stimuli and motor outputs. Where in the brain stimuli and outputs are connected in the brain has remained a difficult question, but of some brain regions it is now clear that they are important in conditioning. For classical conditioning, the cerebellum plays a major role (Steinmetz, 1998; Thompson, 1990). The basal ganglia are known to be involved in operant conditioning, and to code for rewards (Lauwereyns, Watanabe, Coe, & Hikosaka, 2002; Pagnoni, Zink, Montague, & Berns, 2002; Peoples, Uzwiak, Gee, & West, 1997). In fear conditioning, the amygdala has been identified as a central structure (LeDoux, 1996).

These structures are not the only important ones. Conditioning can occur to very complex stimuli as well as to simple ones, such as an entire environment in contextual fear conditioning (Fanselow, 2000). It is unlikely that the structures enumerated above do the necessary stimulus processing, as they are small relative to the size of the whole brain. More likely, they take advantage of processing elsewhere in the brain, using processed, high-level representations as their input.

Indeed, there are several lines of evidence that point to the importance of neocortical processing areas for conditioning. Primary sensory cortices have, for example, been found to reorganize themselves under influence of contingencies in conditioning tasks (Recanzone, Schreiner, & Merzenich, 1993). In addition, the hippocampal region seems to have a role in incremental learning, as is suggested by electrophysiological recordings during conditioning tasks: during eye blink conditioning, hippocampal neural activity develops in response to a conditioned stimulus, and its shape and duration mimic the conditioned response (Berger & Thompson, 1978a).

Hippocampal involvement in incremental learning is also evident from the effects of hippocampal lesions on conditioning paradigms. Although, for example, classical conditioning itself is not slowed by hippocampal damage (Schmaltz & Theios, 1972; Shohamy, Allen, & Gluck, 2000), animals with lesions in the hippocampal region will in many variations of the basic paradigm perform differently than normal animals. Animals with lesions to the hippocampus and overlying cortex do not show the decreased rates of responding typically seen in intact ones when conditioned stimuli are presented in a different context than the one in which conditioning took place (Penick & Solomon, 1991). They also do not show the temporary dip in responding that occurs in intact animals when a CS predictive of a US is suddenly accompanied by another stimulus (Allen, Padilla, Myers, & Gluck, 2002).

The broader hippocampal region (including adjacent parahippocampal cortices and the subiculum) seems to be important for latent inhibition and learned irrelevance. In latent inhibition, the CS is presented many times to the animal in the environment in which training will take place, prior to training. When training starts and the CS is followed by a US, the animal is severely slowed down acquiring the CS-US association (Lubow & Moore, 1959). In learned irrelevance, the animal is first subjected to many explicitly uncorrelated presentations of the stimulus and the US. When in the training phase the CS is reliably followed by the US, learning the CS-US association is slowed relative to a control condition which did not receive the uncorrelated pre-exposure (Mackintosh, 1973). Both phenomena, which tap complex learning about environmental contingencies, are abolished after hippocampal region ablations

INCREMENTAL LEARNING & EPISODIC MEMORY

(Ackil, Mellgren, Halgren, & Frommer, 1969; Allen, Chelius, & Gluck, 2002; Han, Gallagher, & Holland, 1995).

The Gluck and Myers model of incremental learning

One theory of incremental learning, the Gluck and Myers (1993) model, has accounted for these effects of hippocampal lesions by modeling the broader hippocampal region as a predictive autoencoder, which interacts with a simpler module representing the cerebellum (see Figure 1). The task of this autoencoder is to predict its environment, including the presence or absence of all stimuli, including CSs, contextual stimuli, and a US, if any. This leads to the formation of representations that the autoencoder can use to predict an upcoming US. With the predictive encoder intact, the model reproduces standard phenomena in the conditioning literature. With the autoencoder lesioned, it reproduces the above-mentioned effects of hippocampal lesions on conditioning.

Figure 1 about here

Several features of autoencoders help to make it a good model of the hippocampal role in classical and operant conditioning. Autoencoders will compress features that consistently occur together (i.e., have the same predictive value). Features that do not occur together will, on the other hand, be differentiated. This explains the model's performance in paradigms in which contextual discrimination plays a role: a stimulus A may be rewarding in context X, but not in context Y, while it is the other way round for a stimulus B. A simple neural network such as a perceptron cannot solve this nonlinear problem, because its response to A and X together is by necessity the sum of its responses to A and X separately, and neither on its own will predict the reward. The model solves this problem by developing a separate representation for the compound of A and X, and this compound can then predict occurrence of the reward (Myers & Gluck, 1994).

The tendency of autoencoders to compress features that occur together also explains paradigms such as latent inhibition and learned irrelevance. In these paradigms stimulus and context become bound into a single representation during exposure, in which neither predicts anything other than itself. This works against subsequent learning to respond in the presence of the CS but not the context alone.

Another feature of autoencoders is that the amount of learning taking place is a function of prediction error. If the error is large, as it will be at the onset of learning, learning takes big steps. If, after training, the error is small (i.e., the autoencoder can predict its environment and the US), learning will be slow. This explains blocking, a paradigm in which a CS is paired with a US at a moment that the US is already well predicted. In blocking, an animal first goes through many trials in which a stimulus A is paired with a US. When the animal reliably makes a CR to the stimulus, stimulus A is presented in combination with a novel stimulus B, and that compound is still followed by the US. When, after many such trials stimulus B is presented on its own, animals typically emit CRs only at very low rates (Kamin, 1969). No stimulus B – CR link has been made, even though stimulus B was predictive of a US on all trials in which it was presented. In the Gluck and Myers (1993) model, this is explained by the fact that the animal makes no errors during training on the compound, as it already emits CRs to stimulus A. Without errors to be corrected, no learning occurs in the cerebellar

network in which CSs and CRs are connected (this explanation essentially follows from the Rescorla-Wagner rule discussed in a later section).

Limitations of the model

The hippocampal role in many incremental learning paradigms can thus be captured by the mathematics of error correction learning that underlies predictive autoencoders. Three features stand out: that the model can form representations at the right level of complexity (i.e., it forms compound or differentiated representations as needed), that its hippocampal network only learns swiftly when a stimulus has not already been incorporated into the context, and that it learns new CS-CR associations only when a US is not yet well-predicted.

Nevertheless, there are clear grounds for improvement on the model. One is that the model does not clearly map to brain anatomy. For example, the model has one module for the hippocampal region, and can thus not differentiate the role of the hippocampus proper from that of the surrounding areas. Another is that the model does not provide any account of episodic memory in the hippocampus. A third is that error correction learning may indeed occur in the cerebellum, but it has not been shown to take place in the hippocampus. The necessary prediction error signal has never been identified there, and neither has a response commensurate with error reduction. Below, we will present a model that takes over many features of the Gluck and Myers (1993) model, but seeks to remedy the limitations just noted.

Incremental and episodic in one model.

At the heart of our model is a welding of two theories: the Gluck and Myers model, and a generic version of an episodic memory model. From the Gluck and Myers model, it takes an architecture in which a hippocampal memory system interacts with more output-oriented brain structures. But the autoencoder hippocampal submodel is replaced by a multilayer model capable of forming episodic memories (Meeter et al., 2002; Talamini et al., in press). This simple model of episodic memory, as shown in Figure 2, contains three layers in which stimuli and, more broadly, the environment of the organism are represented. An input layer, modeling the neocortex, codes for stimuli and context features. A second layer stands for the parahippocampal region: the perirhinal, entorhinal and postrhinal / parahippocampal cortices. This layer has integrated representations, with some parts coding mostly for context features and some mostly for stimuli, but all parts also getting input of the other kind. The third layer stands for the hippocampus proper, with nodes representing dentate granule cells and/or pyramidal cells in Ammon's horn. This hippocampal layer forms a compact code for the whole situation in which the organism finds itself, for which we use the term 'ensemble' (Murnane, Phelps, & Malmberg, 1999). Such representations form the basis of episodic memory.

Figure 2 about here

To simulate incremental learning tasks, the model has to contain modules coding for the outputs of memory. For classical conditioning, the cerebellum is most relevant, but for other tasks of incremental learning one would have to include output regions for rewards and operant behaviors (basal ganglia) and for fear responses (amygdala). The cerebellar circuit, the only output structure implemented here, is a simplified version of the cerebellar network

in the Gluck and Myers (1993) model. All three representational layers project to the output modules. These connections allow the output modules to attach behavioral significance to simple and complex representations of the same set of stimuli, thereby allowing stimulus configurations to have different associations than the constituent stimuli on their own.

Multilevel representations

This may explain one feature of the Gluck and Myers (1993) model, the formation of compound and single representations. Instead of representations at the right level of complexity being slowly formed through error-correction learning, they may be formed automatically in episodic memory on the first presentation of the stimuli. Conditioning may then proceed by associating events such as the occurrence of a US with episodic representations at the right level of complexity.

To give an example, a simple stimulus such as a tone is coded for in the neocortical layer of the model. This same tone, mixed in with contextual elements, is coded for in the parahippocampal layer. The hippocampal layer has one compound representation that stands for the ensemble of all available cues – i.e., the situation in which the animal finds itself. An outcome can now be connected to the neocortical representation, as in simple conditioning to a tone independent of the context in which the tone occurs (connection labeled ‘a’ in Figure 2). It can also be connected to a parahippocampal representation, which in this model would be equivalent to conditioning to a tone-in-a-context (connection labeled ‘b’ in Figure 2). Finally, it can be connected to a hippocampal representation, which would be equivalent to conditioning to a whole situation (connection labeled ‘c’ in Figure 2). In most cases all three associations may develop at the same time. It may also be, however, that over the course of many learning trials, low level features are associated with other outcomes than compounds of the same stimuli. This would be the case in contextually modulated conditioning.

Effects of outcome predictability

In addition to its ability to form representations of the right complexity, the second feature mentioned above of the Gluck and Myers model is its ability to learn at the right time. Speeded acquisition of a conditioned response (CR) or an operant response occurs only when an unpredicted outcome is presented together with a novel stimulus. When a stimulus loses its novelty before it is combined with an unpredicted outcome, both the model and experimental animals suffer from learned irrelevance or latent inhibition. When the novel stimulus occurs after the outcome has lost its unpredictability, blocking will impede learning.

The effects of outcome predictability on learning are captured very well by the Rescorla-Wagner rule of associative learning (Rescorla & Wagner, 1972). This rule assumes that the amount of learning taking place at any one trial is a function of how well the outcome is, on that trial, predicted by all cues taken together (see appendix for a mathematical formulation). When an outcome paired with a CS is already predicted by other cues, no learning to the CS will take place, and blocking ensues. The circuitry in the cerebellum has been shown to indeed implement the Rescorla-Wagner rule (Kim, Krupa, & Thompson, 1998; Thompson & Gluck, 1991), and it is part of both the Gluck and Myers (1993) model and the cerebellar network of our current model. The current model thus explains blocking in the same way as the Gluck and Myers model does.

In classical conditioning, learning is thus explicitly dependent on prediction failure. Although not directly relevant to the current work, it is important to note that a dependence of learning

INCREMENTAL LEARNING & EPISODIC MEMORY

on outcome predictability is also plausible for kinds of incremental learning involving different brain regions:

- In the basal ganglia, long-term potentiation, a candidate mechanism for long-term memory, is enhanced by dopamine (Thomas & Malenka, 2003), a neuromodulator. Electrophysiological data suggests that dopamine release reflects the unpredictability of current rewards (Schultz, 1998, 2002). These two facts together imply that learning in the basal ganglia is fast when unpredicted rewards occur, and slow once a reward is well predicted. Operant conditioning, thought to rely at least partly on the basal ganglia (Lauwereyns et al., 2002; Pagnoni et al., 2002; Peoples et al., 1997), may thus be governed by the same rules as classical conditioning.
- In fear conditioning, Fanselow (1998) has argued that the interplay between opioid receptors and fear elicitors in the amygdala may be equivalent to the Rescorla-Wagner rule. In particular, if an animal has learned a fear response to a CS, this CS may elicit the release of endogenous opioids, which dampens the effect of the negative reinforcer. In this way, the reinforcing effect of a fear elicitor may decrease during learning, as it does in the Rescorla-Wagner rule.

Stimulus novelty, SOP and familiarity

As described above, stimulus novelty also affects incremental learning. These effects have been formalized in other theories of associative learning (Mackintosh, 1975; Pearce & Hall, 1980), one of which is Sometimes Opponent Processes or SOP (Wagner, 1981). Although SOP has many subtleties, the mechanism by which novelty affects incremental learning is fairly straightforward (see also Donegan, Gluck, & Thompson, 1989). The theory assumes that stimulus representations can be in one of two states (three, if one includes absolute quiescence). When a stimulus is novel to the animal, it will upon presentation bring its representation into a very active state A1. Later, when the stimulus has been presented often in a certain context, the animal will come to expect the stimulus in that context. In this case, when the stimulus is presented, its representation enters in a lower state of activity, A2. SOP assumes that a stimulus representation can be associated with a US only when it is in state A1, not in A2. In latent inhibition, the stimulus is first presented alone for a number of trials. At the end of this pre-exposure, the stimulus will elicit only A2 activity. Since learning does not take place in A2, associating the stimulus with the US will be retarded, relative to a non-exposed control condition in which the stimulus is novel and hence in A1. SOP also predicts that latent inhibition is reduced or eliminated by a change in context between pre-exposure and training phases: the new context will not be associated with the stimulus, and will therefore not bring it into state A2 on presentation. This release from latent inhibition following context shift has indeed been found (Mackintosh, Kaye, & Bennett, 1991; Symonds & Hall, 1995).

SOP can be contrasted with a development in the episodic memory field. Since the seventies, many researchers in the field of episodic recognition memory have argued that recognition judgments need not be based on retrieval of the item to-be recognized, but can also be based on a more fuzzy feeling that the item matches old memories (e.g., Atkinson & Juola, 1974; Humphreys, Bain, & Pike, 1989; Jacoby, 1991; Mandler, 1980; Yonelinas, 2002). These two kinds of memory are often referred to as recollection and familiarity. Recollection is the retrieval of a particular memory, as when an item must be reproduced in recall tasks. Familiarity is a signal computed through a comparison of a memory probe with all memories in the memory store. During a recognition task, the probe is thought of as eliciting a strong familiarity signal if it resembles some or many stored memories. Formalized in computational

models of recognition, the contribution of the familiarity signal can explain many dissociations between recognition and recall (Humphreys et al., 1989; Norman & O'Reilly, 2003; Raaijmakers & Shiffrin, 1992).

Lesion studies (Aggleton & Brown, 1999), electrophysiological evidence (Zhu et al., 1997), and imaging data (Ranganath et al., 2004) have all implicated the perirhinal cortex in familiarity. In the models, familiarity is modeled as an increase in a scalar signal coming out of the memory system. When neural responses in the perirhinal cortex to novel and familiar stimuli are directly compared, however, results seem to contradict this idea. Instead of being larger for familiar stimuli, neural responses to a stimulus decrease in the perirhinal cortex with increasing familiarity (Li et al., 1993; Xiang & Brown, 1998). Although these results (see Figure 3 for example) contradict many theories of familiarity, they are exactly what one would expect from SOP: a state of high activity at the first presentations, with less activity being elicited by an oft-repeated stimulus. Here, we will suggest that the familiarity effect, the decreased parahippocampal response to familiar stimuli, is what causes the effects of stimulus novelty on the speed of conditioning. This leaves open how this familiarity effect is caused.

Figure 3 about here

A theory of familiarity

The decrease in firing for familiar stimuli effects may result from the interaction between a few uncontroversial neural mechanisms. The basic idea is that stimulus representations may always partly reflect the latest context in which they were seen. This will make them responsive to that context on itself, which will in turn lead to less responsiveness to the stimulus if presented in the same context due to adaptation and accommodation.

Figure 4 shows how a parahippocampal node may come to represent context as well as its preferred stimulus. The input to the parahippocampal gyrus during an experiment consists of a phasic input when a stimulus is briefly presented, and of more or less continuous stimuli that together represent context (e.g., relatively constant features of the physical environment). We will assume that each parahippocampal node receives one strong connection from the input it codes for, and weak connections from other inputs. Although this is a gross simplification, the underlying idea that parahippocampal cells have preferred stimuli is well-supported (see below).

In panel *a* of Figure 4, connections for one parahippocampal node coding for a stimulus are drawn in. This node receives a strong connection from a lower-layer node coding for its preferred stimulus, but other inputs, together forming the context, also reach it. If the preferred stimulus is presented, the node will fire strongly, and the weak connections from contextual inputs will be strengthened through LTP. These connections may now be strong enough to make the node fire at a low rate when its preferred stimulus is absent (panel *b*). This in turn will lead to adaptation in the node. When the preferred stimulus is presented again in that context, the node will be in a less responsive state due to the built-up adaptation (panel *c*). It will therefore respond at a lower rate to it than when the stimulus was first presented. Exactly this pattern been found in perirhinal and entorhinal neurons (Li et al., 1993; Xiang & Brown, 1998).

On a side note, firing decrements have not yet been observed in the parahippocampal cortex. This seems at odds with our inclusion of the parahippocampal cortex in the substrate of our

parahippocampal layer. The parahippocampal cortex relays spatial layout information, among other things, to the hippocampus (Bohbot, Allen, & Nadel, 2000; Vann, Brown, Erichsen, & Aggleton, 2000), and we felt it was important to include this source of hippocampal inputs in our model. To our knowledge it has not been investigated whether parahippocampal neurons show a familiarity firing decrement for preferred stimuli (this is the case in our model, but it plays no role in the current simulations).

Figure 4 about here

Summary of the model

In summary, we propose a new model of brain regions involved in learning and memory. With it we hope to elucidate phenomena from the literature of episodic memory and classical conditioning. It contains three layers that form representations, and that stand for the neocortex, the parahippocampal gyrus, and the hippocampus proper (see Figure 2). Of the output regions in Figure 2 only the cerebellum is implemented for this paper. The basic assumptions of the theory behind our model, as presented above, are listed below. A comparison of these assumptions with those of other models is left to the discussion.

1. The neocortex, parahippocampal gyrus and hippocampus form a hierarchy in which episodic representations of increasing complexity are formed.
2. Stimulus representations at different levels of complexity may acquire different or even opposite associations.
3. Learning in the cerebellum is governed by the Rescorla-Wagner rule. This explains why incremental learning is modulated by outcome predictability.
4. Incremental learning is modulated by stimulus novelty, as suggested by SOP, because firing to stimuli in the parahippocampal gyrus decreases with increasing familiarity.
5. Firing to familiar stimuli decreases in the parahippocampal gyrus because stimulus representations become responsive to the context in which stimuli are presented.

Steps one to four do not depend on the fifth assumption being correct. They require only that there is a difference between the parahippocampal activity elicited by novel and familiar stimuli, and this has already been observed, as described above. Assumption 5 is one plausible way in which this could occur (other theories of the familiarity firing decrement are reviewed in the general discussion).

Model implementation

The five assumptions above have been implemented as a computational model, which we will here describe at a conceptual level. Formulas and technical aspects of the implementation are relegated to the appendix.

The three representational modules are implemented as three layers of linear input integrators with a continuous firing rate. The architecture is loosely based on a model of episodic

memory (Meeter et al., 2002; Talamini et al., in press); feedback connections existing in that model are omitted here and a more explicit representation of neuronal activity is chosen. The cerebellar network is taken from previous work (Gluck, Allen, Myers, & Thompson, 2001; Gluck & Myers, 1993); in its basis, it is a simple perceptron in which an input layer codes for stimuli and an output layer codes for a CR. Weights between these layers are slowly modified with the Rescorla-Wagner rule in order to predict a target output (equivalent to the US).

Firing rate of the representational nodes is a thresholded, linear function of cell activity, which is itself equal to summed inputs multiplied by a factor accounting for adaptation. Adaptation is a function of previous node firing rate, while inputs to a node consist of node-specific excitation and undifferentiated feedforward inhibition. Feedforward inhibition is a linear function of the total activity in the layer below, excitation a weighted sum of the outputs of its nodes. Weights on the excitatory connections can change through learning, governed by a variant of Hebb's rule often used in competitive learning, the Oja rule (Oja, 1982). This rule models both long-term potentiation (LTP) and heterosynaptic long-term depression (LTD): weights are strengthened when pre- and postsynaptic nodes are both active, and weakened when the postsynaptic node is active while the presynaptic node is not. Time is discrete, with iterations standing for half a second.

Connections from the cortical to the parahippocampal layer are not uniform. In the brain, distinct pathways carry distinct kinds of inputs to the hippocampus, but these streams become more and more intertwined as they make their way through the parahippocampal gyrus (Witter, Wouterlood, Naber, & Van Haeften, 2000). In the model, this is simplified so that each parahippocampal node receives a strong, topological connection from a corresponding neocortical node, and weak, distributed connections from other neocortical nodes. Functionally, this scheme implies that a parahippocampal node has a preferred stimulus it reacts strongly to, but also becomes weakly active in response to other inputs, meaning that context can influence representations in the parahippocampal layer. Indeed, parahippocampal primary neurons are known to be stimulus-selective in their responses, but to also be contextually modulated (Dusek & Eichenbaum, 1997; Suzuki, Miller, & Desimone, 1997).

Connections from the parahippocampal layer to the hippocampal layer are dense and fanning. They model the perforant path projections from entorhinal cortex to dentate gyrus and hippocampal field CA3, which have such characteristics (Witter et al., 2000). This arrangement allows the hippocampal layer to form ensemble representations, coding for the combination of all stimuli in the parahippocampal layer. Compared with connections from the cortical to the parahippocampal layer, initial weights are larger in this connection, but this is balanced by the absence of the strong, topological connections that are present in the lower connections. Feedforward inhibition is proportional to initial weight strength, and thus also is higher in the parahippocampal-to-hippocampal connections than in the cortex-to-parahippocampal connections.

The three representational layers send projections to each one third of the input layer of the cerebellar network. Connections from nodes in the representational layers to the cerebellar input layer are all one-to-one. Auditory cortical areas are known to project to the cerebellar cortex via the basilar pontine nuclei (Aitkin & Boyd, 1978). There is less evidence with regard to mesocortical afferents. Anatomically, no direct projections from the hippocampal region to the cerebellum have been found. However, indirect, polysynaptic connections have been suggested to exist (Berger, Weikart, Bassett, & Orr, 1986). For example, a retrograde tracing study in which viruses were injected in the cerebellum found that both the hippocampus and the entorhinal cortex projected indirectly to the cerebellum (Kaufman, Mustari, Miselis, & Perachio, 1996). Moreover, stimulation in field CA1 of the dorsal hippocampus leads to a response in the cerebellum (Yu, Wang, & Chen, 1989). One indirect pathway may be from the subiculum to the ventromedial hypothalamus (Kohler, 1990), a

region that itself projects to the cerebellum (Haines, May, & Dietrichs, 1990). Another possibility is that projections go via entorhinal and other cortical areas. (As hippocampal outputs are processed in entorhinal deep layers while feedforward processing takes place in the superficial layers of that structure, the assumption of independence of parahippocampal and hippocampal projections to the cerebellum would still be tenable). Nevertheless, until the existence of such indirect connections has received stronger anatomical or neurophysiological support, it remains a critical assumption of the model.

The effects of septal cholinergic innervation of the hippocampus are modeled by including an extra learning phase whenever one of two situations applies. First, following several models of episodic memory (Hasselmo & Wyble, 1997; Meeter et al., 2004), septal activity is assumed whenever parahippocampal inputs fail to elicit activity in the hippocampus. Such lack of activity indicates a novel situation. The mechanism assumed by these models, disinhibition of the septum if firing in the hippocampus is low, is consistent with data indicating an inhibitory effect of hippocampus on the septum (Dragoi, Carpi, Recce, Csicsvari, & Buzsaki, 1999; McLennan & Miller, 1974). The other situation in which we assume septal activity, following Rokers et al. (2002), is in the presence of an unpredicted US. Indeed, it has been found that a US tends to activate the medial septum at the beginning of training, while this activity tapers off when conditioned responses start appearing (Berger & Thompson, 1978b). In both situations eliciting septal activity, a subset of nodes with the largest inputs undergoes LTP and LTD with a larger learning parameter, modeling one of the physiological effects of acetylcholine (Hasselmo, 1999).

Reported results are all averages from ten replications.

Results: episodic memory

The primary goal of the model is to show how the hippocampal region can have a role in both episodic memory and incremental learning. First, we will discuss the ways in which the present model addresses findings in episodic memory.

Above, a distinction commonly made by theories of episodic memory was described, namely that between familiarity and recollection. Norman and O'Reilly (2003) showed that many characteristics of the two processes can be explained by assuming that they are differently responsive to input overlap; see Yonelinas (2001) for a presentation of a similar idea presented in a more formal framework. They equated familiarity with a signal that varies continuously with the overlap of probes with previously stored memories, and recollection with a process that only yields an output when the probe closely resembles a stored memory. The left panel of Figure 5 shows how the strength of the two signals depends on the overlap between the probe and previously stored items. Familiarity follows input overlap nearly linearly, while recollection only delivers a strong signal at large input overlaps. The right panel of the figure shows how this may explain the relative usefulness of familiarity and recollection with different kinds of lures. In recognition, targets (studied items) have to be discriminated from unstudied lures. These lures can be similar to targets or dissimilar, with similar lures overlapping in more features with targets than dissimilar ones. The panel shows how the two signals differ for targets, similar and dissimilar lures. As can be seen, targets can be separated quite easily from similar lures with both a familiarity and a recollection signal, but only in recollection is the difference between a target and a similar lure large. (On a side note, Norman and O'Reilly show that very similar lures lead to retrieval errors, or "false memories", making recollection unreliable when such lures are used).

Figure 5 about here

Here, we will first show how our model produces both signals. We show then how signal strength in our model depends on input overlap in the same way as in the Norman and O'Reilly (2003) model.

Familiarity

To test the theory of familiarity presented in the introduction, we presented a stimulus A twice to the parahippocampal layer in one context. We then compared the response to the second presentation of A with the response to a second, novel stimulus B. Each stimulus was represented by 1 cortical node, active for 10 iterations (i.e., 5 seconds), and was preceded by 20 iterations of only context presentation. As in all further simulations, 15 active cortical nodes represented context, and active cortical nodes had an output of 0.5, inactive ones of 0.

Figure 6a plots the response of the parahippocampal node maximally responsive to first presentation of A. It shows that the response to a repetition of A is indeed lower than that to a novel stimulus B, replicating the basic finding (Li et al., 1993; Xiang & Brown, 1998). Note that at the second presentation, the parahippocampal node has started responding to context, which drives up its adaptation and thus lowers its response to the stimulus.

Figure 6 about here

The response decrement for repetitions occurs under a wide range of parameter sets, but both its strength and the time course of its induction are determined by parameter values. Three important factors are weight change during one presentation (less change will lead to a slower induction of the familiarity effect), adaptation constants (stronger adaptation will lead to a stronger effect), and the weight distribution at the onset of the simulation. Such factors may partly explain the variability in neuronal responses to familiarity reported by Xiang and Brown (1998). These authors analyzed responses as a function of two variables: repetition within sessions, and repetition in different sessions. The analysis led to four types of neurons: **Novelty neurons** were those that reduced their responses to both within-session and between-session repetition. They fired strongly to novel patterns, but not to any repetition. **Familiarity neurons** were those that were influenced by between-session repetitions, but showed no significant effect of within-session repetition. These neurons seemed to gradually decrease their firing with continued repetition. **Recency neurons** were influenced by within-session repetition but not by between-session repetition. They decreased their firing when a stimulus had been presented recently, but not when it had been presented in earlier sessions. Finally, **visually responsive neurons** were not influenced by repetition of any kind.

The results shown in Figure 6a represent a novelty neuron, one that fires differently on the first presentation of a stimulus as compared to all later presentations. It was produced using a high parahippocampal learning rate (0.1). In the data of Xiang and Brown (1998), responses to familiar patterns were only around 30-50% of that to the original presentation. With a learning rate of 0.1 the ratio of responses to novel and familiar stimuli was around 50% (this drops to below 25% with even higher learning rates).

Figure 6b shows the results with the low learning rate used in the remaining simulations (0.01). Here, the effect takes shape over a number of trials. This would make the node akin to

INCREMENTAL LEARNING & EPISODIC MEMORY

Xiang and Brown's (1998) familiarity neurons, neurons whose firing to familiar patterns (i.e., ones seen many times) is lowered as compared to that to unfamiliar stimuli, but not necessarily showing a decrement in firing on the first repetition of a pattern.

Simulating Xiang and Brown's (1998), recency neurons, would necessitate the assumption of weight decay mechanisms, which are not present in the model. If weights decayed between sessions, neurons would decrement firing in response to repetitions within sessions but not across sessions. Xiang and Brown's (1998) novelty neurons also show stronger firing decrements when a repetition is within a session than when it is in another session, suggesting that such decay may exist for all recorded neurons, but is parametrically stronger in recency neurons.

The fourth kind of neuron, visually responsive neurons whose responses were no different for repetitions than for first presentations, were actually the most common in the data set of Xiang and Brown (1998). Very low learning rates could produce such neurons, but they may also be thought of as the neurons coding that receive no or too few inputs from afferent cells coding for contextual elements to be influenced by context. In summary, variations in wiring, weight decay and learning rate may produce the four kinds of neurons found by Xiang and Brown (1998).

One further thing to note about Figure 6a is the transient nature of much of the firing elicited by the stimulus. The strong firing in the first time step that the stimulus is present quickly returns to lower levels through adaptation. This is highly realistic: even in situations in which there is a sustained neural response to a stimulus, the sustained response is usually preceded by a stronger transient signal at stimulus onset (e.g., Lamme, Rodriguez-Rodriguez, & Spekreijse, 1999; Tsujimoto & Sawaguchi, 2004; Xiang & Brown, 1998). As a consequence processing in our model is biased towards stimulus onsets; even though context representations are always larger than stimulus representations, responses elicited by stimuli at onset are in the same range or stronger than the summated response to all context elements. This will be important in our simulations of classical conditioning. As only novel, phasically present stimuli elicit a large response in the parahippocampal layer, the system will have a built-in, automatic bias to associate a US with such stimuli instead of with constant context elements.

Recollection

A second way in which to query episodic memory is recollection or retrieval. This is usually modeled as pattern completion, occurring after a pattern has been strengthened by Hebbian learning (Hasselmo & Wyble, 1997; McClelland & Goddard, 1996; Meeter et al., 2004; Norman & O'Reilly, 2003; Talamini et al., in press). To test whether our model would show similar behavior, we performed a simulation in which a pattern was presented for five time steps in a context (again represented by 15 cortical nodes). Later in the simulation, either a degraded version of the pattern (with 40% of features set to 0), or an entirely novel pattern, was presented in the same context. To allow for partial patterns, patterns in our recollection simulations consisted of 5 cortical nodes, whereas they consist of 1 cortical node in other simulations. This does not change the results, as a pattern of one cortical node also elicits the creation of a new hippocampal pattern.

Figure 7a plots parahippocampal and hippocampal activity in the model, averaged over the whole layers as representations form over time. In the first time steps, a hippocampal representation is formed of the context. In the very first time step no hippocampal activation results from the large input; the random, initial weights are not strong enough for

hippocampal nodes to overcome feedforward inhibition (weaker feedforward inhibition and strong input connections from the cortical layer to the parahippocampal layer precludes such silencing by feedforward inhibition in the parahippocampal layer). In vivo electrophysiological recordings have indeed shown a period of several hundreds of milliseconds of severely dampened firing in CA3 and CA1 following presentation of a novel stimulus (Vinogradova, Kitchigina, & Zenchenko, 1998). The silence in the hippocampal layer triggers septal activation, and the formation of a hippocampal pattern to represent its input. In the next time step, a hippocampal pattern has been formed and becomes active. Thereafter, both parahippocampal and hippocampal activation decrease due to adaptation, until a plateau is reached at around the tenth time step. At time step 16, the pattern is presented (black bar in left panel of Figure 7a), leading to a strong response in the parahippocampal layer and a second interruption of firing in the hippocampal layer. Again, the random weights are not strong enough for hippocampal nodes to overcome the feedforward inhibition resulting from the new inputs. The interruption once more triggers septal activation, and a hippocampal pattern is now formed to represent the item in its context.

Figure 7 about here

When the pattern is presented again in degraded form (see right panel in Figure 7a), it immediately activates its hippocampal representation. A novel pattern does not activate it; instead, it does not elicit any activity for one time step. In the next time step, this new pattern has itself been stored under influence of septal activation. In the first time step in which a pattern is presented, hippocampal activity thus distinguishes sharply between old patterns, which elicit activity even in degraded form, and novel patterns, which do not. This contrast can form the basis for recollection-based recognition in ways explored by other models of episodic memory.

As a measure of pattern retrieval, we use the correlation over all hippocampal nodes between the hippocampal representations of the first pattern and either the degraded or the novel pattern. As can be seen in Figure 7b, this correlation was very large for the degraded pattern, but negative for the novel pattern. The full hippocampal pattern was produced with partial cues, which constitutes pattern completion in the hippocampus. Strengthened connections between nodes in the parahippocampal and hippocampal patterns underlies this pattern completion.

Comparison of recollection and familiarity

To investigate the relation between input overlap and parahippocampal and hippocampal activity, we performed a simulation in which the overlap between a first and a second stimulus was systematically varied. A stimulus represented by 5 neocortical nodes was presented once for 5 time steps in a standard context. After a delay, a second 5-node stimulus was presented that overlapped in 0, 1, 2, 3, 4 or 5 nodes with the first stimulus. The 0 overlap condition was equivalent to the novel pattern condition in the previous simulation.

Figure 7 showed that parahippocampal and hippocampal responses to stimulus repetition are opposite: whereas parahippocampal nodes fire more strongly to a novel pattern, the hippocampal layer responds more strongly to an old pattern. The same pattern emerges when pattern overlap is systematically varied. With decreasing overlap, and thus increasing pattern novelty, parahippocampal responses increase while hippocampal responses decrease (see

Figure 8a). To ease comparison, we rescaled both signals so that both were maximal at 100% overlap (i.e., straight repetition of pattern 1).¹ As the inset in Figure 8a shows, our model reproduces, though in an exaggerated way, the responses of familiarity and recollection in the Norman and O'Reilly (2003) model to input overlap (see Figure 5a). Parahippocampal response to the second pattern varies linearly with input overlap, while hippocampal response depended in a very nonlinear way on input overlap. As our model reproduces the central distinction between familiarity and recollection in the Norman and O'Reilly (2003) model, it also reproduces their results on similar and dissimilar lures (see Figure 8b).

Figure 8 about here

The magnification of small differences between inputs to large differences between hippocampal patterns also explains another feature of the model. Although context alone and context plus stimulus overlap in most features (namely in all context features), the addition of the stimulus nevertheless causes a whole new pattern to be formed (see Figure 7). This is the case even when a stimulus consists of a single node (because of adaptation to context, at first presentation the stimulus will produce a large parahippocampal activity compared to the context; see Figure 6).

Discussion

Although our simulations of episodic memory are not very elaborate, they can be seen as a proof of concept: they show that our architecture is sufficient to model basic episodic memory tasks, based on either recall / recollection or familiarity. In the case of recollection this is not surprising, as the model conforms in many ways to the common denominator of computational models of episodic memory. The idea that parahippocampal areas compute a familiarity signal is also widespread (Aggleton & Brown, 1999; Bogacz, Brown, & Giraud-Carrier, 2001; Norman & O'Reilly, 2003), but this signal is implemented here in a novel way: as a decrease of parahippocampal activity caused by adaptation to the current context. Several predictions that follow from this novel implementation are given in the general discussion.

Familiarity in our proposal is context-specific - to a lesser extent, this is also the case in other implementations of familiarity (see discussion). In some theories, however, familiarity is thought of as explicitly context-free (e.g., Mandler, 1980). The “butcher on the bus” phenomenon, the familiar face that we cannot place because we see it outside the normal context, suggests familiarity is context-independent. However, there may also be butchers on the bus who we miss completely because they evoke no familiarity outside of their usual context.

In controlled studies, simple context manipulations sometimes influence human recognition memory and sometimes do not. A recent meta-analysis concluded, however, that a global context shift (e.g., a room change) has as much effect on recognition as it has on recall scores (Smith & Vela, 2001). This does not imply that the two hypothetical underlying processes, familiarity and recollection, are both context-dependent. The two most pertinent studies differ radically in their conclusions. Macken (2002) concluded on the basis of the remember/know

¹ The transformation for the parahippocampal response was: $(r_{100}-r_x) / (r_{100}-r_0)$, where r_x is the response to a particular level of overlap, r_{100} is the response to 100% overlap (i.e., to straight repetition), and r_0 is the response to 0% overlap (i.e., to a wholly new pattern). For the hippocampal response it was: $(r_x-r_{100}) / (r_0-r_{100})$.

paradigm that all context effects in recognition derived from the contributions of recollection estimates, and that familiarity was context-insensitive. McKenzie and Tiberghien (2004) found the exact opposite result with a process dissociation procedure: small context effects in recollection, large ones in familiarity estimates. A factor that may play a role in these conflicting findings is the timing of context presentations. In our model, context works by slightly activating associated stimuli in the interval before they are presented. When context and item information are attended to at the same time, the effect of context should thus be attenuated or eliminated. Macken (2002), who found familiarity to be unaffected by context, presented context and item information at the same time, which suggests they were also attended to at the same time. In McKenzie and Tiberghien's (2004) study, added context elements were presented before the item was, and opposite conclusions were reached.

Although stimulus timing may thus be a factor, it is clear that the question of the context-specificity of familiarity is still open. The question whether all familiarity relies of the parahippocampal context is also not wholly solved. It would be odd if, for faces, specialized face processing systems (Farah, Wilson, Drain, & Tanaka, 1998; Kendrick, da Costa, Leigh, Hinton, & Peirce, 2001) do not also play a role in computing familiarity.

One aspect of context effects in recognition is that a context change tends to result not only in a lower hit rate, but also in a lower false alarm rate (e.g., Macken, 2002). At first sight, this is paradoxical, as it implies that a context shift lowers familiarity for lures never presented in that context. It is, however, uncontroversial if we think of items as represented by multiple features. If lures overlap in some features with targets (studied items), neurons coding for these features will show a familiarity firing decrement in the old context but not in a new one. Therefore, firing would be lowered for lures sharing features with targets in the old context, but not in new ones. This implies that context effects should interact with target-lure similarity, in that a context change should affect false alarm rates more for similar lures than for non-similar lures. This stands as a prediction of the model.

Results: classical conditioning

The previous section has shown how the model implements both recall/recollection, and familiarity. Recall / recollection is based on the reactivation of compound representations formed at acquisition, familiarity on the decreased responding to familiar stimuli. The episodic memory simulations pave the way for our work on classical conditioning. The familiarity effect modeled in the parahippocampal layer implements the central idea of SOP, described in the introduction, that only novel stimuli elicit a state suitable for fast acquisition of responses to them. The formation of episodic memories in the hippocampal layer allows for conditioning to occur to complex, compound representations.

We will first describe our simulations of basic conditioning. The role of different representations in generating simple conditioned responses will be investigated both through inspection of individual layer contributions, and through lesion studies. We will then investigate several paradigms from the classical conditioning literature. Our aim in these simulations is twofold: to show that the intact model can reproduce findings in the literature, and to show that lesions in the model have the same consequences for performance as they have in experimental animals.

Basic conditioning

In our simulations of classical conditioning, no attempt was made to quantitatively fit animal data. Instead, we assumed a generic setup in which an animal is placed in a context, is allowed to familiarize itself with it, and is then subjected to 99 trials of conditioning to a CS followed by a US. Context consisted again of 15 continuously active cortical nodes, with both the stimulus and the US being represented by 1 active cortical node. A CS had a duration of one time step (1/2 second), and was immediately followed by a US of also one time step. Between each presentation of the stimulus were 20 iterations (10 seconds) in which only the context was active. As our measure of conditioning we took the output of the cerebellar network, which we assume to be monotonically related to the likelihood of a CR.

Figure 9, left panel, shows how the output of the cerebellum in the time step after a CS rises with increasing numbers of presentations of the CS. It also shows the decomposition of the output into contributions from inputs from the three layers. These were computed by summing the contribution to cerebellar output of cerebellar nodes receiving input from a particular layer. As can be seen, the hippocampal and parahippocampal layers contribute most to the generation of cerebellar output, with the cortical contribution rising slowly throughout training. The fact that parahippocampal and hippocampal contributions rise faster than those of the cortical layer, is the result of strong responses to novel stimuli: whereas in cortex the CS is just one of many active elements, it generates a much larger signal in the parahippocampal layer because it is novel and only phasically active. Since hippocampal output is a function of parahippocampal input, the CS elicits strong activity also in this layer. The CS thus generates a stronger signal in parahippocampal and hippocampal layers, making it easier for the cerebellum to form a CS-US connection using inputs from these layers. Later in learning, parahippocampal responses to the more and more familiar CS decrease. The advantage of the parahippocampal CS representation over its cortical predecessor thus disappears, which results in that later in learning, the cortical CS representation accrues more strength as a CR elicitor in the cerebellum (see later portion of learning curve).

The hippocampal contribution does not come right at the onset of training. The right panel of Figure 9 shows performance on the first ten trials. The hippocampus only starts contributing to cerebellar output after a few trials. In those trials, a hippocampal representation of the CS-context combination has to be formed and then strengthened under influence of the septum.

Figure 10 shows cerebellar output to the context alone. Throughout the whole training episode, the model correctly generates no CRs to context alone. Underneath this nonresponse lies a revealing pattern: through the training, hippocampal and to a lesser degree parahippocampal inputs come to drive the cerebellum to incorrectly emit a CR to context alone, a tendency that is only inhibited through negative contributions from the cortical layer. This reflects pattern completion: the context alone is part of the pattern of context and stimulus, and will therefore weakly activate the hippocampal representation of the context-stimulus ensemble even in the absence of the stimulus. Moreover, it will weakly activate parahippocampal stimulus representations. The context alone thus comes to prefigure the stimuli that regularly appear in it, and therefore activate CS-US associations that were formed within it.² Although that may sound like an intuitive finding, it leads to the counterintuitive prediction that individual elements making up the context come to have negative predictive value during training. They should thus function as mild inhibitors when presented in a

² Activation of the context-CS ensemble in the context-alone situation will also weaken its association with the US, as no US follows context-alone presentations. This counteracts further learning of the CS-US learning, which is another reason for the stagnant parahippocampal and hippocampal contributions to the CR.

different context. This could be tested in a experiment in which animals are classically conditioned in a context X. Salient elements of that context could be removed, and later placed back. Removal of the elements should lead to increased responding, returning them to the context to a drop in responding.

Figure 9 about here

Figure 10 about here

Effects of lesions

To investigate the roles of different layers, we repeated the above simulation with one or more layers lesioned. Three lesions were investigated. In one, the hippocampal layer was removed, modeling a selective hippocampal lesion in animals. In the second, both the hippocampal and the parahippocampal layers were removed, modeling a broader hippocampal region lesion. Finally, we investigated selective lesions of the medial septum by not applying the extra learning phase under influence of acetylcholine. In this simulation, only base-rate learning occurred in the hippocampus. We made the assumption that after a lesion of a representational layer, remaining inputs become stronger through the compensatory processes generally seen when an area or neuron is partly deafferented: synapses from remaining connections grow larger and more numerous (Robertson & Murre, 1999). This was modeled by multiplying remaining inputs to the cerebellum with a factor representing the proportion of inputs lost (i.e., if only the hippocampal layer was lesioned, remaining inputs were multiplied by 1.5; if both the hippocampal and parahippocampal layers were lesioned, remaining inputs were multiplied by 3).

Under this assumption, CS-CR associations were formed at the same speed after lesions of the hippocampal and/or parahippocampal layers as in the intact model (Figure 11a). Acquisition of the CR was slowed substantially only after a septal lesion. The Gluck and Myers model makes the same predictions, and for the same reason. With a removal of the hippocampal region, cortical inputs still support the formation of CRs to simple stimuli. After a septal lesion, no new representations are formed in the hippocampal layer, but hippocampal outputs are still projected to the cerebellum. In both the Gluck and Myers (1993) model and in the current one, no hippocampal input is better than a dysfunctional hippocampal region input. This pattern is consistent with findings from rabbit eye blink conditioning (Figure 11b): acquisition of an eye blink response to a simple cue is not impaired after hippocampal region lesions (Schmaltz & Theios, 1972) or selective lesions of the hippocampus (Allen, Padilla, Myers et al., 2002), but is slowed after medial septal lesions (Allen, Padilla, & Gluck, 2002). In our model, the results depend critically on the assumption of synaptic compensation. Although the weight multiplication factor can be somewhat below the level of total compensation, if it were dramatically lower or if compensation did not occur in the cerebellum, lesions to the hippocampal or parahippocampal layer would affect the speed of conditioning.

The lesions discussed above were made before training. Lesions can also be made during or after training. What the model would predict in such cases can be gleaned from Figure 9, by

comparing the full cerebellar response to what it would be if one or more components were missing. The hippocampal and parahippocampal layers play a large role in generating cerebellar output early in training, with the cortical layer contributing substantially only later in training. If the hippocampus is lesioned early in training, performance will thus suffer to a large extent, with the effect becoming gradually smaller if the lesion is made later in training. In a similar vein, Figure 9 shows that a lesion encompassing parahippocampal regions as well as the hippocampus proper will affect performance in all stages of training. Our model thus predicts no effects of both types of lesions when they are made before training, but effects of varying strength when they are made during or after training. Although we know of no data showing these effects for classical conditioning, they have been found in fear conditioning, in which lesions sufficient to cause retrograde amnesia (i.e., loss of trained responses) hardly affect the acquisition of new responses (Anagnostaras, Gale, & Fanselow, 2001).

Figure 11 about here

Sensitivity to context in the intact and lesioned model

If after conditioning in one context, the response is tested in another, a decrement in performance is often found (Penick & Solomon, 1991). It is as if a change in context removes part of the cues for the CR. This effect is not universally found, however. Reviewing many experiments, Myers and Gluck (1994) suggested that the amount of training is an important determinant, with response decrements likely if the context is changed early in training, but less likely if the context is changed after extensive training.

Figure 12a shows how cerebellar output changes in the model when context is changed after a certain number of trials. The effect of context change is relatively large after 30 trials but becomes smaller later in training, as was argued to be the case in experimental animals (Myers & Gluck, 1994). The Gluck and Myers (1993) model explained this pattern by a tuning of representations. Early in training, the CS and context are part of a single representation; later in training, representations for the rewarded CS and the unrewarded context become more and more separate. Our model produces the same pattern for a similar reason: early in training, the hippocampal layer plays a larger role in generating cerebellar output than later in training (see Figure 9). As the hippocampal layer codes for stimulus-context ensembles, CRs thus depend on representations that are highly context-laden early in training, but less so later in training. This leads to gradually smaller context effects with more training. Both the current and the Gluck and Myers (1993) model predict that overtraining should eventually abolish context effects.

Figure 12b shows the effects of a context change after 30 trials for lesioned models. A decrement is only seen in the intact model. There was no context shift decrement after a hippocampal region lesion. Penick and Solomon (1991) showed that context change effects were indeed abolished following hippocampal lesions (Figure 12c). In the model, responses are even strengthened by a context change after such a lesion. This results from the fact that familiarity decreases responses in a context-dependent way. A change of context will therefore increase the parahippocampal response to the CS. As cerebellar output is a linear function of its inputs, this increase in parahippocampal response will translate into a larger cerebellar output, resulting in more, not less, responding in the new context. In several data sets, responses are indeed larger in the novel context than in the old one after hippocampal lesions. This has been found in rabbit eye blink conditioning (Figure 12c), but also in rat

operant conditioning (Honey & Good, 1993). It is unclear whether this effect was reliable in these data sets, so it stands as a prediction of the model. The model further predicts disruption of context effects by septal lesions, which also remains to be tested empirically.

Figure 12 about here

Blocking

Blocking refers to the phenomenon that an animal will not learn a connection between a stimulus and an outcome (say, a US) that is already well predicted. As in that of Gluck and Myers (1993), the model's explanation for blocking relies on the presence of a prediction error term in the Rescorla-Wager rule describing learning in the cerebellum (see appendix). A well-predicted US does not elicit learning in the cerebellum, because there is no prediction error to drive learning. As this effect depends on the cerebellum, lesions in the hippocampal region should not affect it. This is indeed what was found in the model. After 99 trials in which stimulus A was paired with a US, 99 trials were given in which stimulus A was joined by another stimulus B. In the last trial, only stimulus B was presented. A strong blocking effect was found in all conditions, consisting of no responding to B alone. No lesion abolished the blocking effect (see Figure 13a). This replicates experimental findings (Figure 13b) in that neither hippocampal lesions (Allen, Padilla, Myers et al., 2002) nor medial septum lesions (Baxter, Gallagher, & Holland, 1999) abolish blocking.

Hippocampal lesions do have a more subtle effect, however. In intact animals, introduction of the novel cue causes a temporary drop in the rate of responding (Allen, Padilla, Myers et al., 2002; Rokers et al., 2002). This "novel cue effect" (see Figure 13d) is abolished after hippocampal lesions (Allen, Padilla, Myers et al., 2002). Our model reproduces these findings, as the novel cue causes a new hippocampal representation to be formed, disrupting older representations that underlie part of the cerebellar output. Figure 13c plots how output on the last trial before the introduction of the new stimulus, with output on the ten trials after introduction. Consistent with the rabbit data, a decrement in response was found in the intact model, but not in the model with hippocampal lesions. The predictions for the broader hippocampal region and medial septal lesions remain to be tested empirically.

Figure 13 about here

Learned irrelevance and latent inhibition

Learned irrelevance and latent inhibition are produced, in the model, through the effects of familiarity. A novel cue elicits strong firing in the parahippocampal region, which makes it easy to attach significance to the cue. Familiar stimuli elicit much weaker firing, rendering conditioning to such stimuli more difficult. This then results in learned irrelevance or latent inhibition.

In our latent inhibition simulation, stimulus A was first presented for 99 trials on its own. Subsequently, A was paired with a US for another 99 trials. In such situations responses are

acquired more slowly than in a control condition with no stimulus pre-exposure. The model reproduced this finding: in the latent inhibition condition, a response developed more slowly than in a control condition in which the model was given only context exposure prior to the 99 CS-US pairings (Figure 14). At this point, the CS in the control condition was novel, and thus elicited large responses in the parahippocampal nodes. This made it easy for the cerebellar layer to attach predictive value to the CS. In the latent inhibition condition the CS was familiar during the CS-US pairings, and thus did not elicit strong parahippocampal responses anymore.

In the learned irrelevance condition, a CS and US were both presented 99 times during a first training phase, but in an unpaired fashion (each US was presented somewhere in the 20 time step interval in between two CS presentations). In a second phase the CS was paired with the US for 99 trials. This produced slower learning than in the control condition (same as in the latent inhibition case; Figure 14). Again, predictive value is more easily attached to novel stimuli than to stimuli with which the animal has become familiar.

Although latent inhibition is a robust phenomenon, it is generally eliminated by a change in context between the two learning phases (Mackintosh et al., 1991; Symonds & Hall, 1995). As was already discussed, SOP predicts this finding, and as the current model is in a sense an implementation of SOP, it is no surprise that the model also reproduces the finding. In our simulation, we presented a stimulus for 99 trials in context 1, and then paired it for 99 trials with the US in context 2. Learning now proceeded at the same speed as in the non-exposed control condition (Figure 14). A context change restores parahippocampal firing to the level of a novel stimulus, and therefore allows conditioning to the stimulus to occur at normal speed.

In additional simulations, we investigated the sensitivity to lesions of latent inhibition, and also of the effect of context change. The results were unambiguous: because in our model latent inhibition is a result of the lower parahippocampal firing in response to familiar stimuli, only lesions involving the parahippocampal layer interfere with latent inhibition (Figure 15). This was indeed found in rabbit eyeblink conditioning: parahippocampal (specifically, entorhinal cortex) lesions did interfere with latent inhibition, while lesions restricted to the hippocampus proper did not (Shohamy et al., 2000). Although septal lesions have not yet been investigated in this paradigm, learned irrelevance was still present with systemic injections with scopolamine, which disrupt septohippocampal cholinergic projections (Moore, Goodell, & Solomon, 1976). The sparing of latent inhibition by medial septal lesions stands as a prediction of the model.

Figure 14 about here

Figure 15 about here

Discussion

Although at first blush episodic memory and incremental learning seem to pose contradictory demands, the computational model presented in this paper shows how the two may in fact be

INCREMENTAL LEARNING & EPISODIC MEMORY

complementary. The properties of episodic representations in the neocortex, parahippocampal gyrus and hippocampus proper may explain phenomena in classical conditioning. The model reproduces the effects of hippocampal, septal, and broad hippocampal-region lesions on classical conditioning and its contextual modulation. This is done while respecting gross brain anatomy, and taking into account many findings from neurophysiology and neuropharmacology.

What the model accounts for or can account for

Many theories of episodic memory distinguish between two outputs of memory: familiarity and recollection. We have described how each output may arise from different parts of the hippocampal region, as has been argued previously by others (Aggleton & Brown, 1999; Norman & O'Reilly, 2003; Yonelinas, 2002). A familiarity signal may arise from parahippocampal cortices, and a recollection / recall output may be provided by the hippocampus proper. Our model diverges from others only in how the familiarity signal arises: as a decrease in signal due to increasing contextual sensitivity.

Our simulations of episodic memory are essentially a proof of principle. We showed that familiarity and recollection have the same characteristics in our model as they have in other models. Familiarity is sensitive to the overlap between the memory probe and stored memories. Recollection is sensitive to strong overlap with a single memory, and will generate complete memories to degraded cues. These characteristics help to simulate a large number of findings from episodic memory, as has been shown by others (Humphreys et al., 1989; Norman & O'Reilly, 2003; Talamini et al., in press; Yonelinas, 2001). Our current model would have to be scaled up to entertain such simulations, however, as it is currently too small to support storage of lists of items.

The characteristics of familiarity and recollection were used to simulate several paradigms of incremental learning. Ensemble representations formed in the hippocampal layer could explain the drop in CRs when a CS is presented in a new context. It could also explain the "novel cue" effect in blocking. The drop in parahippocampal response to stimuli with increasing familiarity reproduced learned irrelevance and latent inhibition. The fact that parahippocampal representations include contextual information produces a release from learned irrelevance and latent inhibition with context change.

We did not attempt to quantitatively fit precise experiments. This had several reasons. One was that because of the simplicity of the cerebellar model, its output is not directly comparable to the number of CR's emitted by the animal. Direct fitting would have required a mapping function, which due to a lack of data would have been arbitrary. Instead, we chose to do all simulations with one fixed set of parameter values and present our results as qualitative fits of known effects. This does mean, however, that effect sizes in simulations cannot be compared to those in experiments – only their direction can be validly compared.

The results are quite robust for changes in these parameter values. One parameter that does have a strong impact is the relative size of a context representation as compared to a CS representation. Representing a context by too few nodes can lead to a domination of hippocampal representations by the CS, eliminating context effects. Too large context representations make learning slow, as it takes the cerebellum more time to pick up the CS 'signal' in the context 'noise'. Other important parameters are those governing the balance between LTP and LTD in the model. If LTP is too strong as compared to LTD 'runaway synaptic modification' (Hasselmo, 1994) occurs, pre-empting the creation of useful hippocampal representations. Too strong LTD would lead to continuously dwindling weights,

which is not a very realistic model of memory. Parameters with mostly quantitative effects (i.e., on the size but not the direction of effects) are the relative balance between layer inputs to the cerebellum, and the parameters governing learning, septal activity, and adaptation (not all sets produced strong familiarity effects).

The model lends itself quite naturally to extensions into operant conditioning, similar to how Myers and Gluck (1996) expanded the Gluck and Myers (1993) model to apply to operant data by incorporating multiple output nodes and making reinforcement contingent on output responding. More paradigms of episodic memory could also be incorporated, as could conditioned fear. This does not mean that successful extension is guaranteed. Notably, our model currently does not include a realistic source of variability, which precludes simulation of all characteristics of distributions of responses. Our model currently exhibits only some sampling variability in hippocampal representations, which is why ten replications of each simulation were sufficient. Future editions of the model will need to address this point (see Norman & O'Reilly, 2003 for an excellent discussion of the issue of variability in neural networks).

What the model does not account for

There are also data that the model does not account for, and aspects of the model that are clear simplifications. The rich literature on extinction is a case in point. Many complex phenomena in this literature implicate the hippocampus (Frohardt, Guarraci, & Bouton, 2000), and our model seems ideally placed to make a contribution to this literature. However, preliminary simulations showed that it already fails to provide an adequate explanation of extinction at a low level. It often takes much longer to extinguish a CR to a stimulus than it takes to learn it, as if the system errs on the side of caution. However, preliminary simulation showed that our model extinguishes a response at approximately the speed at which it acquires them, as do other associative models of conditioning. This is because the Rescorla-Wagner rule at the core of our cerebellar model decreases the CS-CR association after an omission (CR with no US) in the same way as it increases that association after a commission (no CR before the US). However, recent research has suggested that the cerebellar mechanisms behind acquisition and extinction of CR's are different (Krupa & Thompson, 2003).

Another area in which improvements could be made is the time course at which hippocampal responses occur during conditioning (Berger & Thompson, 1978a). It takes more trials to develop strong responses in the rat than it takes here, and they start off as responses to the US, which here they do not. It would be possible to fit this data, as we have incorporated the US in the input to the hippocampus. In the current setup that would not serve any purpose; a model on a more fine time scale might be needed to give meaningful interpretations to the data. Although this leaves something to be explained, it does not challenge the model as it stands.

The model does not make a contribution to a long-standing discussion in the episodic memory literature, namely whether episodic memories are consolidated in the neocortex or not (for review, see Meeter & Murre, 2004). Both the view that they are consolidated in the neocortex (McClelland, McNaughton, & O'Reilly, 1995; Meeter & Murre, in press; Squire & Alvarez, 1995; Squire, Cohen, & Nadel, 1984) and the view that they are not (Nadel & Moscovitch, 1997; Nadel, Samsonovitch, Ryan, & Moscovitch, 2000) could be reconciled with the current model.

An area to which the model also does not make a contribution is the timing of responses, and the influence of timing of conditioned and unconditioned stimuli on the acquisition of responses. Here, we chose for a CS lasting one time step and a US that immediately follows

the CS, which makes the model not dissimilar to trial-based models with added context-only trials (e.g., Gluck & Myers, 1993). In contrast with trial-based models, however, the explicit representation of time makes our model essentially capable of being extended to other timing regimes. Two additions would be necessary. Delay conditioning could be simulated by making the elicited CR part of the input of the cerebellar layer, as has been shown previously (Gluck et al., 2001). The second addition would be a loop between the parahippocampal and hippocampal layers, as has been implemented in models of episodic memory (e.g., Talamini et al., in press). This would allow representations to remain active in both regions through reverberations, as has indeed been observed between the hippocampus and entorhinal cortex (Iijima et al., 1996). Reverberatory representations could be used to form associations in trace conditioning tasks, in which CS and US are not coactive. Indeed, some evidence has pointed to the hippocampus as being necessary in delay conditioning (Clark, Manns, & Squire, 2001; Clark & Squire, 1999).

Relation to previous models

The model is a variant of associative memory theories, and as such it has many of the strengths and weaknesses of these theories (e.g., in the field of extinction). It has borrowed heavily from two associative theories in particular, that of Rescorla and Wagner (1972) and SOP (Wagner, 1981), bringing the latter from an abstract level to a neurobiologically testable level (see also Donegan et al., 1989). Central elements of both theories are implemented in the current model.

The current model replaces and builds on the Gluck and Myers model of cortico-hippocampal function in classical and operant conditioning (Gluck & Myers, 1993, 2001; Myers & Gluck, 1996; Myers & Gluck, 1994; Myers, Gluck, & Granger, 1995), and its extensions to septo-hippocampal interactions (Myers et al., 1996; Rokers et al., 2002) and cerebellar function in eyeblink conditioning (Gluck et al., 2001; Gluck & Thompson, 1987). The implementation of the cerebellum and septohippocampal system were adapted from these models. The implementation of the hippocampal region was changed –there as an autoassociator, here as a hierarchy of representational layers. This allowed us to circumvent three limitations of the previous model: that it was anatomically imprecise (remedied only to a degree in the current model), that it placed unobserved error-correcting learning in the hippocampus, and that it did not create room for episodic memories in that structure. It also means that at a mechanistic level, the two models differ substantially. An example is the speed at which representations are formed. Whereas in the previous model representations were slowly tuned under influence of behavioral contingencies, here we take the episodic memory perspective in which the bulk of hippocampal learning occurs at the first confrontation with a novel situation or stimulus, and occurs in an unsupervised fashion.

At a more functional level, however, many explanations of the current model are very similar to those of the earlier framework. As an example, we consider the proposed mechanisms behind latent inhibition. In the Gluck and Myers (1993) model, this phenomenon is explained through a hippocampal region-mediated compression of the CS with the context during the first phase of training, in which the CS and US are presented in an uncoupled fashion. This makes the task of learning a CS-US association in the second phase of training hard, as the model cannot differentiate between context alone and context plus CS. In the current model, compression of CS and context into one representation occurs in the hippocampus, but it is not necessary for latent inhibition. During the first phase of learning, the CS stops eliciting a strong response in its habitual context (the familiarity effect), and this is assumed to underlie latent inhibition. Although the details of both explanations differ, they bear family

resemblance: both assume that the CS has more or less "faded" into the background context during the first phase of learning. At a more abstract level, the two models thus deliver similar explanations for the findings in the conditioning literature. They also share many predictions, such as a declining influence of context on classical conditioning.

As was already mentioned, the model is also similar to many neural network models of episodic memory. The structure of its representational layers was loosely based on a model of episodic memory (Meeter et al., 2002; Talamini et al., in press), and its behavior in the episodic memory simulations is similar to that of other episodic memory models (Hasselmo & Wyble, 1997; McClelland & Goddard, 1996; Meeter et al., 2004; Norman & O'Reilly, 2003). Our assumptions about septal activity are also widely shared by episodic memory models (Hasselmo & Wyble, 1997; Meeter & Murre, in press; Meeter et al., 2004).

In many respects, the current model thus merely pulls together threads that were evident in earlier models of the hippocampus. In at least two aspects it is different from those other models:

- In its explicit modeling of the interaction between the cerebellum and episodic representations stored in the hippocampal region.
- In how sensitivity to context in parahippocampal representations leads to lower responses to familiar stimuli.

Comparison with theories of familiarity

The proposed explanation of the familiarity effect is not the only one, however. It has been suggested that a few perirhinal neurons coding strongly for the stimulus show increased firing rates, which in turn inhibit other neurons (Norman & O'Reilly, 2003). Since there would be more neurons with decreased responses, these would be the ones observed in studies of familiarity neurons. The observed decreases may also reflect neurons that lose out in the self-organization phase in which a new representation is formed for a novel stimulus (Sohal & Hasselmo, 2000). Both proposals suggest that whereas firing is decreased in many neurons, it is increased in "winners", those neurons that really code for the stimulus. However, although Xiang and Brown (1998) recorded from hundreds of neurons, none showed an increase in firing rate.

One model suggests one reason why such increases might not, in fact, occur. In the model of Bogacz et al. (2001), "winners" inhibit themselves via feedback loops. Although firing is increased in the first milliseconds after a stimulus appears, it is quickly extinguished, leading to an overall decrease in firing during repetitions even in these winners. With some parameter settings, the Norman and O'Reilly (2003) model could behave in a similar way. The initial increase in firing frequency was not found by Xiang and Brown (1998); they note that from the first spike on firing rates are lower for familiar stimuli. However, it is possible that noise in response times causes the initial increase to be invisible, as the graphs of Xiang and Brown (1998) are averaged over many trials.

The same authors propose another model, in which familiarity neurons activated strongly by a novel stimulus undergo LTD instead of LTP. This so-called **anti-Hebbian model** (Bogacz & Brown, 2002, 2003) provides an elegant explanation for the reduced firing to familiar stimuli, because 'winners' firing strongly at the first presentation of a stimulus will, through LTD, limit their firing during repetitions. The model does not depend on inhibition for the decrease, and does not predict firing increases in any group of neurons. However, it presupposes that LTD occurs between highly active pre- and postsynaptic neurons, which is counter to the

standard situations in which LTD occurs and has yet not been observed in neurons (Bogacz & Brown, 2003).

Another advantage of the anti-Hebbian model is its large theoretical capacity: with the same number of neurons, a model based on anti-Hebbian learning could store many more patterns than some models based on Hebbian learning (Bogacz & Brown, 2003). Models based on Hebbian learning increase their capacity, however, by including homosynaptic LTD— LTD observed when a presynaptic neuron is active but the postsynaptic neuron is not (Bogacz & Brown, 2002). This lowers the correlation between neurons not involved in coding for the same pattern, and a mathematical analysis has shown that an on average positive correlation between firing in different perirhinal nodes is the main limiting factor in the capacity of a perirhinal memory store (Bogacz & Brown, 2003). Such LTD was not needed in the current model, but it could be included without changes to the simulations.

In summary, all published models of the familiarity effect, as well as the current proposal, leave something to be proven. The proposals based on inhibition or reorganization would seem to require either an increase in the firing frequency of some familiarity neurons, either continuously or immediately after a stimulus has been presented. This has not been observed, but is possible that averaging over trials obscures the increase. The current proposal and anti-Hebbian learning may be easier to reconcile with existing data, but both make strong claims that have so far not been tested. The current proposal suggests that the familiarity effect should diminish with context change, and even disappear with a total context change. The anti-Hebbian model suggests that LTD should dominate LTP in the perirhinal cortex.

As already stated in the introduction, the current implementation of familiarity could be replaced by other theories without endangering most simulation results. There is one caveat, however. One effect, latent inhibition, does depend on an aspect of the theory, namely on the context-dependence of familiarity. As discussed in the introduction, SOP explains the context-dependence of latent inhibition by assuming that the context of learning brings representations in a state of low activity. The equation of familiarity with this state in SOP is thus only possible if familiarity is context-dependent. Familiarity is indeed context-dependent in our implementation, and it also would be in the theory of Norman and O'Reilly (2003), in the anti-Hebbian learning model, and in the theory of Sohal and Hasselmo (2000). In all three, context features would presumably be part of the input to parahippocampal / MTLC / perirhinal modules, and thus be partly responsible for the match between study items and test probes.

Comparison with Complementary Learning Systems

The Norman and O'Reilly (2003) model of episodic memory is an instantiation of a broader framework, dubbed Complementary Learning Systems (McClelland et al., 1995; O'Reilly & Rudy, 2000). Within the same framework, O'Reilly and Rudy (2001) presented simulations both of tasks that could be seen as episodic, and of paradigms from incremental learning. Their focus in the latter is on operant and fear conditioning. They simulate four sets of findings: nonlinear discrimination problems, contextual fear conditioning, transitivity problems, and context-sensitive responding. In the latter three sets, the simulations rely on pattern completion in their hippocampal system as the causal mechanism. In particular, O'Reilly and Rudy show how the hippocampus could build up a compound representation of context, which could become associated for example with a shock in the contextual fear paradigm. Since pattern completion is also essential to the functioning of our hippocampal layer, our model would simulate these sets in the same way as their model (although it would require a theory of the formation of context representations that would take into account the

role of stimulus novelty in our model). In the last case, that of context-sensitive responding, this is obvious from our simulations of the same phenomenon in classical conditioning. The first set, of nonlinear discrimination problems, would be treated fundamentally differently in our model than in their framework.

Nonlinear discrimination problems are those in which patterns of stimuli must be discriminated by the animal that are not linearly separable. An example is negative patterning, a paradigm in which two stimuli, A and B, predict a reward, but their combination AB does not. In most experiments, experimental animals will take a long time to correctly respond to A and B, but not to AB. Such tasks require a separate representation for the combination AB, which can then be associated with different consequences than the individual stimuli A and B. Such compound representations would naturally be formed in our hippocampal layer, and indeed many models of hippocampal functioning would predict that the hippocampus is necessary for negative patterning and similar nonlinear discrimination problems (Sutherland & Rudy, 1989). However, experiments have shown that animals with hippocampal lesions can still perform on such tasks (Alvarado & Rudy, 1995; Davidson, McKernan, & Jarrard, 1993), although they are often slowed compared to normal animals (Alvarado & Rudy, 1995; Rudy & Sutherland, 1995). This seems to require compound representations (i.e., for the compound AB) outside of the hippocampus. In O'Reilly and Rudy's (2001) model, the cortex can form these representations through error-correction learning. In our model, such representations would automatically be formed, because stimuli occurring simultaneously form each other's context. Parahippocampal nodes coding for A would thus become "loaded" with B just as they become loaded with context. If some such nodes received inputs coding for B and others did not, their differential responses to the AB compound could form the basis for nonlinear discriminations such as negative patterning.

The different ways in which the two model deal with negative patterning point to a major difference, namely in learning rules. The O'Reilly and Rudy (2001) model assumes that two forms of learning, Hebbian and error-correction learning, are present in both the cortical and the hippocampal memory systems. Hebbian learning underlies performance in episodic tasks, while error correction is important for the incremental learning simulations. This seems unparsimonious, but a justification could be that the difference reflects task demands: error signals are important in incremental learning tasks, but not in episodic memory tasks. Our main argument not to follow their assumption is that error signals and error-driven learning have not been observed in the hippocampus or in cortex (though see O'Reilly, 1996 for a defense of error-driven learning in cortical areas). This contrasts with the cerebellum, where it has been convincingly demonstrated (Kim et al., 1998; Thompson & Gluck, 1991), and with the basal ganglia and amygdala that could operate under similar regimes (see introduction). The result of this difference is that in the O'Reilly and Rudy model, representations are tuned by error correction, while in our model representations are strictly episodic, and tuning occurs in the connections from the representational layers to output systems such as the cerebellum.

There are, however, also many features in which the Complementary Learning Systems framework overlaps with the one laid out in this paper. For example, the hippocampus in both frameworks automatically forms rapid compound representations of the situation, and these representations explain context-dependence in incremental learning tasks. At a more fundamental level, the two agree in that the same memory representations may underlie very different forms of learning in animals (as expressed by classical, operant and fear conditioning), and that there is a fundamental continuity between human and animal memory tasks. They also overlap in what is a fundamentally associative view of incremental learning: in both models, representations of stimuli are associated with outputs that drive responses.

Comparison with comparator theories

The strongest contrast is probably between our model, firmly in the associative tradition, and comparator theories (Gallistel, 1990; Gallistel & Gibbon, 2000; Gibbon & Balsam, 1981; Miller & Matzel, 1988). The latter claim that conditioning is not a matter of forming associations between stimuli and responses, but that it involves estimations as to when an emitted response will be rewarded. Gallistel and Gibbon (2000) have made a convincing claim that current associative theories are incomplete, and have not yet resolved certain internal inconsistencies within the associative framework. We agree that there are phenomena, especially in the timing domain, for which associative theories do not fully account. We hope to be able to extend our model in that direction.

Nevertheless, there are also areas in which associative theories offer better explanations than timing theories, notably in describing context effects. For example, it is not obvious within a framework such as Rate Estimation Theory (Gallistel, 1990; Gallistel & Gibbon, 2000) why context change would eliminate latent inhibition, as the expectations generated by stimuli are not affected by context in these models. The biggest drawback of comparator theories, however, is that most are entirely functional, with few guidelines as to how they might be implemented in the brain. Associative theories have an advantage here, as shown perhaps most clearly by findings on the cerebellum and the Rescorla-Wagner rule (Kim et al., 1998; Thompson, 1990). Perhaps, when comparator theories are made biologically more plausible and associative theories are brought to bear on timing issues, they will converge onto a common framework.

Untested predictions

Many predictions follow from the model as it stands. As a favor to our future falsifiers, we will end with a simple enumeration of these. We start with those that follow exclusively from our theory of familiarity, and from its dependence on context:

1. If there is a change in context between presentations of a stimulus, no firing decrement should be observed in parahippocampal areas.
2. Familiarity estimates should show context dependence.
3. Familiarity should depend on context being present before the stimulus is presented.
4. A context change should affect false alarm rates more for similar lures than for non-similar lures.
5. Decrements in firing to familiar stimuli in parahippocampal areas should be preceded by increases in background firing rate in the same cell, reflecting on stronger sensitivity to context alone.

Other predictions do not depend on our theory of familiarity being correct, although some rely on the fact that familiar stimuli elicit smaller parahippocampal responses (as has been found). Predictions 6, 7 and 8 are shared with the Gluck and Myers (1993) model; the others are not.

6. A lesion to the hippocampal region after training will affect even simple conditioned responses that can be learned normally if the lesion is made before training.

INCREMENTAL LEARNING & EPISODIC MEMORY

7. The effect of a lesion to the hippocampal region on a conditioned response will be larger during early phases of training (when hippocampal representations play a relatively large role, see Figure 9) than later in training.
8. Overtraining will abolish or severely limit context change effects in classical conditioning.
9. Medial septal lesions will abolish context change effects in simple conditioning.
10. Context effects in simple conditioning will be in the opposite direction in animals with lesions restricted to the hippocampus proper: they will show an increment in responding after context change.
11. Individual elements making up the context come to have negative predictive value during training. They should thus function as inhibitors when presented in a different context.

Although vigorous testing of these and other predictions of our model will almost certainly prove it wrong in details (and perhaps wrong in its entirety), we believe that the model highlights an advantage of computational modeling: that knowledge from different task domains and brain areas can be integrated into one coherent framework, and lead to testable predictions. With more and more knowledge being generated about the brain, we believe that such an integrating function will become more and more essential.

References

- Ackil, J., Mellgren, R., Halgren, C., & Frommer, G. (1969). Effects of CS preexposures on avoidance learning in rats with hippocampal lesions. *Journal of Comparative and Physiological Psychology*, *69*(4), 739-747.
- Aggleton, J.P., & Brown, M. W. (1999). Episodic memory, amnesia, and the hippocampal-anterior thalamic axis. *Behavioral and Brain Sciences*, *22*, 425-489.
- Aitkin, L., & Boyd, J. C. (1978). Acoustic input to the lateral pontine nuclei. *Hearing Research*, *1*, 17-77.
- Allen, M. T., Chelius, L., & Gluck, M. A. (2002). Selective entorhinal and nonselective cortical-hippocampal region lesions, but not selective hippocampal lesions, disrupt learned irrelevance in rabbit eyeblink conditioning. *Cognitive, Affective, and Behavioral Neuroscience*, *2*(3), 214-226.
- Allen, M. T., Padilla, Y., & Gluck, M. A. (2002). Ibotenic acid lesions of the medial septum retard delay eyeblink conditioning in rabbits (*Oryctolagus cuniculus*). *Behavioral Neuroscience*, *116*(4), 733-738.
- Allen, M. T., Padilla, Y., Myers, C. E., & Gluck, M. A. (2002). Selective hippocampal lesions disrupt a novel cue effect but fail to eliminate blocking in rabbit eyeblink conditioning. *Cognitive, Affective, and Behavioral Neuroscience*, *2*, 318-328.
- Alonso, J., Sang U, H., & Amaral, D. (1996). Cholinergic innervation of the primate hippocampal formation: II. Effects of fimbria/fornix transection. *Journal of Comparative Neurology*, *375*, 527-551.

INCREMENTAL LEARNING & EPISODIC MEMORY

- Alvarado, M., & Rudy, J. (1995). A comparison of kainic acid plus colchicine and ibotenic acid-induced hippocampal formation damage on four configural tasks in rats. *Behavioral Neuroscience, 109*(6), 1052-1062.
- Anagnostaras, S. G., Gale, G. D., & Fanselow, M. S. (2001). Hippocampus and contextual fear conditioning: Recent controversies and advances. *Hippocampus*(8-17).
- Atkinson, R. C., & Juola, J. F. (1974). Search and decision processes in recognition memory. In D. H. Krantz, R. C. Atkinson, R. D. Luce & P. Suppes (Eds.), *Contemporary developments in mathematical psychology*. San Francisco: Freeman.
- Baxter, M., Gallagher, M., & Holland, P. (1999). Blocking can occur without losses in attention in rats with selective removal of hippocampal cholinergic input. *Behavioral Neuroscience, 113*(5), 881-890.
- Berger, T. W., & Thompson, R. (1978a). Neuronal plasticity in the limbic system during classical conditioning of the rabbit nictitating membrane response. I. The hippocampus. *Brain Research, 145*(2), 323-346.
- Berger, T. W., & Thompson, R. (1978b). Neuronal plasticity in the limbic system during classical conditioning of the rabbit nictitating membrane response. II: Septum and mammillary bodies. *Brain Research, 156*(2), 293-314.
- Berger, T. W., Weikart, C. L., Bassett, J. L., & Orr, W. B. (1986). Lesions of the retrosplenial cortex produce deficits in reversal learning of the rabbit nictitating membrane response: Implications for potential interactions between hippocampal and cerebellar brain systems. *Behavioral Neuroscience, 100*(802-809).
- Bogacz, R., & Brown, M. W. (2002). The restricted influence of sparseness of coding on the capacity of familiarity discrimination networks. *Network: Computations in Neural Systems, 13*, 457-485.
- Bogacz, R., & Brown, M. W. (2003). Comparison of computational models of familiarity discrimination in the perirhinal cortex. *Hippocampus, 13*, 494-524.
- Bogacz, R., Brown, M. W., & Giraud-Carrier, C. (2001). Model of familiarity discrimination in the perirhinal cortex. *Journal of Computational Neuroscience, 10*, 5-23.
- Bohbot, V. D., Allen, J. J., & Nadel, L. (2000). Memory deficits characterized by patterns of lesions to the hippocampus and parahippocampal cortex. *Annals of the New York Academy of Sciences, 911*, 355-368.
- Cabeza, R., & Nyberg, L. (2000). Imaging cognition II: An empirical review of 275 PET and fMRI studies. *Journal of Cognitive Neuroscience, 12*, 1-47.
- Clark, R., Manns, J., & Squire, L. (2001). Trace and delay eyeblink conditioning: Contrasting phenomena of declarative and nondeclarative memory. *Psychological Science, 12*(4), 304-308.
- Clark, R., & Squire, L. (1999). Human eyeblink classical conditioning: Effects of manipulating awareness of the stimulus contingencies. *Psychological Science, 10*(1), 14-18.
- Davachi, L., Mitchell, J. P., & Wagner, A. D. (2003). Multiple routes to memory: distinct medial temporal lobe processes build item and source memories. *Proceedings of the National Academy of Sciences USA, 100*, 2157-2162.
- Davidson, T., McKernan, M., & Jarrard, L. (1993). Hippocampal lesions do not impair negative patterning: A challenge to configural association theory. *Behavioral Neuroscience, 107*(2), 227-234.
- Donegan, N. H., Gluck, M. A., & Thompson, R. F. (1989). Integrating behavioral and biological models of conditioning. In *Psychology of Learning and Motivation* (Vol. 3, pp. 109-156): Academic Press.
- Dragoi, G., Carpi, M., Recce, M., Csicsvari, J., & Buzsaki, G. (1999). Interactions between hippocampus and medial septum during sharp wave and theta oscillation in the behaving rat. *Journal of Neuroscience, 19*, 6191-6199.

INCREMENTAL LEARNING & EPISODIC MEMORY

- Dusek, J. A., & Eichenbaum, H. (1997). The hippocampus and memory for orderly stimulus relations. *Proc Natl Acad Sci U S A*, *94*(13), 7109-7114.
- Eichenbaum, H. (1992). The hippocampal system and declarative memory in animals. *Journal of Cognitive Neuroscience*, *4*, 217-231.
- Eichenbaum, H. (2000). Hippocampus: mapping or memory? *Current Biology*, *10*(21), R785-R787.
- Fanselow, M. S. (1998). Pavlovian conditioning, negative feedback and blocking: Mechanisms that regulate association formation. *Neuron*, *20*, 625-627.
- Fanselow, M. S. (2000). Contextual fear, gestalt memories, and the hippocampus. *Behavioural Brain Research*, *110*, 73-81.
- Farah, M. J., Wilson, K., Drain, M., & Tanaka, J. (1998). What is "special" about face perception? *Psychological Review*, *105*(1), 482-498.
- Ferbintineau, J., & Shapiro, M. L. (2003). Prospective and retrospective memory coding in the hippocampus. *Neuron*, *40*, 1227-1239.
- Fleidervish, I. A., Friedman, A., & Gutnick, M. J. (1996). Slow inactivation of Na⁺ current and slow cumulative spike adaptation in mouse and guinea-pig neurocortical neurones in slices. *Journal of Physiology (London)*, *493*, 83-97.
- Frohardt, R., Guarraci, F., & Bouton, M. (2000). The effects of neurotoxic hippocampal lesions on two effects of context following fear extinction. *Behavioral Neuroscience*, *114*(2), 227-240.
- Gallistel, C. R. (1990). *The organization of learning*. Cambridge, MA: Bradford Books / MIT Press.
- Gallistel, C. R., & Gibbon, J. (2000). Time, rate, and conditioning. *Psychological Review*, *107*, 289-344.
- Gibbon, J., & Balsam, P. (1981). Spreading association in time. In C. M. Locurto, H. S. Terrace & J. Gibbon (Eds.), *Autoshaping and conditioning theory* (pp. 219-253). New York: Academic Press.
- Gilbert, P., Kesner, R., & Lee, I. (2001). Dissociating hippocampal subregions: A double dissociation between dentate gyrus and CA1. *Hippocampus*, *11*, 626-636.
- Gluck, M. A., Allen, M., Myers, C., & Thompson, R. (2001). Cerebellar substrates for error correction in motor conditioning. *Neurobiology of Learning and Memory*, *76*, 314-341.
- Gluck, M. A., Meeter, M., & Myers, C. E. (2003). Computational models of the hippocampal region: Linking incremental learning and episodic memory. *Trends in Cognitive Sciences*, *7*, 269-276.
- Gluck, M. A., & Myers, C. E. (1993). Hippocampal mediation of stimulus representation: A computational theory. *Hippocampus*, *3*, 491-516.
- Gluck, M. A., & Myers, C. E. (2001). *Gateway to Memory: An Introduction to Neural Network Modeling of the Hippocampus in Learning and Memory*. Cambridge, MA: MIT Press.
- Gluck, M. A., Myers, C. E., & Thompson, R. (1994). A computational model of the cerebellum and motor-reflex conditioning. In S. Zornetzer, J. Davis, C. Lau & T. McKenna (Eds.), *An Introduction to Neural and Electronic Networks* (pp. 91-98). New York: Academic Press.
- Gluck, M. A., & Thompson, R. F. (1987). Modeling the neuron substrate of associative learning and memory: A computational approach. *Psychological Review*, *94*, 176-191.
- Haines, D. E., May, P. J., & Dietrichs, E. (1990). Neuronal connections between the cerebellar nuclei and hypothalamus in *Macaca fascicularis*: cerebello-visceral circuits. *Journal of Comparative Neurology*, *299*, 106-122.
- Han, J. S., Gallagher, M., & Holland, P. (1995). Hippocampal lesions disrupt decrements but not increments in conditioned stimulus processing. *Journal of Neuroscience*, *15*(11), 7323-7329.

INCREMENTAL LEARNING & EPISODIC MEMORY

- Hasselmo, M. E. (1994). Runaway synaptic modification in models of cortex: Implications for Alzheimer's Disease. *Neural Networks*, 7(1), 13-40.
- Hasselmo, M. E. (1999). Neuromodulation: Acetylcholine and memory consolidation. *Trends in Cognitive Sciences*, 3, 351-359.
- Hasselmo, M. E., & Bower, J. M. (1993). Acetylcholine and memory. *Trends in Neuroscience*, 16(6), 218-222.
- Hasselmo, M. E., & Wyble, B. P. (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.
- Honey, R., & Good, M. (1993). Selective hippocampal lesions abolish the contextual specificity of latent inhibition and conditioning. *Behavioral Neuroscience*, 107(1), 23-33.
- Humphreys, M. S., Bain, J. D., & Pike, R. (1989). Different ways to cue a coherent system: a theory for episodic, semantic and procedural tasks. *Psychological Review*, 96(2), 208-233.
- Iijima, T., Witter, M. P., Ichikawa, M., Tominaga, T., Kajiwara, R., & Matsumoto, G. (1996). Entorhinal-hippocampal interactions revealed by real-time imaging. *Science*, 272(5265), 1176-1179.
- Jacoby, L. L. (1991). A process dissociation framework: separating automatic from intentional use of memory. *Journal of Memory and Language*, 30, 513-541.
- Jarrard, L. (1995). What does the hippocampus really do? *Behavioral Brain Research*, 71(1-2), 1-10.
- Kamin, L. (1969). Predictability, surprise, attention and conditioning. In B. C. a. R. Church (Ed.), *Punishment and Aversive Behavior* (pp. 279-296). New York: Appleton-Century-Crofts.
- Kaufman, G. D., Mustari, M. J., Miselis, R. R., & Perachio, A. A. (1996). Transneuronal pathways to the vestibulocerebellum. *Journal of Comparative Neurology*, 370, 501-523.
- Kendrick, K. M., da Costa, A. P., Leigh, A. E., Hinton, M. R., & Peirce, J. W. (2001). Sheep don't forget a face. *Nature*, 414, 165-166.
- Kim, J., Krupa, D., & Thompson, R. (1998). Inhibitory cerebello-olivary projections and blocking effect in classical conditioning. *Science*, 279, 570-573.
- Kohler, C. (1990). Subicular projections to the hypothalamus and brainstem: some novel aspects revealed in the rat by the anterograde Phaseolus vulgaris leucoagglutinin (PHA-L) tracing method. *Progress in Brain Research*, 83, 59-69.
- Krupa, D. J., & Thompson, R. F. (2003). Inhibiting the Expression of a Classically Conditioned Behavior Prevents Its Extinction. *Journal of Neuroscience*, 23, 10577-10584.
- Lamme, V. A. F., Rodriguez-Rodriguez, V., & Spekreijse, H. (1999). Separate processing dynamics for texture elements, boundaries and surfaces in primary visual cortex of the macaque monkey. *Cerebral Cortex*, 9, 406-413.
- Lauwereyns, J., Watanabe, K., Coe, B., & Hikosaka, O. (2002). A neural correlate of response bias in monkey caudate nucleus. *Nature*, 418, 413-417.
- LeDoux, J. (1996). *The emotional brain*. New York: Simon & Schuster.
- Levy, W. B., Colbert, C. M., & Desmond, N. L. (1990). Elemental adaptive processes in neurons and synapses: A statistical/computational perspective. In M. A. Gluck & D. E. Rumelhart (Eds.), *Neuroscience and connectionist theory*. Hillsdale, NJ: Lawrence Erlbaum.
- Li, L., Miller, E. K., & Desimone, R. (1993). The representation of stimulus familiarity in anterior inferior temporal cortex. *Journal of Neurophysiology*, 13, 1918-1929.

INCREMENTAL LEARNING & EPISODIC MEMORY

- Lubow, R., & Moore, A. (1959). Latent Inhibition: The effect of non-reinforced pre-exposure to the conditional stimulus. *Journal of Comparative and Physiological Psychology*, 52(4), 515-519.
- Macken, W. J. (2002). Environmental context and recognition: The role of recollection and familiarity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28, 153-161.
- Mackintosh, N. J. (1973). Stimulus selection: Learning to ignore stimuli that predict no change in reinforcement. In R. H. a. J. Stevenson-Hinde (Ed.), *Constraints on Learning: Limitations and Predispositions* (pp. 75-96). New York: Academic Press.
- Mackintosh, N. J. (1975). A theory of attention: Variations in the associability of stimuli with reinforcement. *Psychological Review*, 82(4), 276-298.
- Mackintosh, N. J., Kaye, H., & Bennett, C. H. (1991). Perceptual learning in flavour aversion conditioning. *Quarterly Journal of Experimental Psychology B*, 43, 297-322.
- Mandler, G. (1980). Recognizing: The judgment of previous occurrence. *Psychological Review*, 87, 252-271.
- Marr, D. (1971). Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society B*, 262, 23-81.
- McClelland, J. L., & Goddard, G. V. (1996). Considerations arising from a complementary learning systems perspective on hippocampus and neocortex. *Hippocampus*, 6, 654-665.
- 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.
- McKenzie, W. A., & Tiberghien, G. (2004). Context effects in recognition memory: The role of familiarity and recollection. *Consciousness and Cognition*, 13, 20-38.
- McLennan, H., & Miller, J. J. (1974). The hippocampal control of neuronal discharges in the septum of the rat. *Journal of Physiology*, 237(3), 607-624.
- Meeter, M., & Murre, J. M. J. (2004). Consolidation of long-term memories: Evidence and alternatives. *Psychological Bulletin*, 130, 843-857.
- Meeter, M., & Murre, J. M. J. (in press). TraceLink: A connectionist model of consolidation and amnesia. *Cognitive Neuropsychology*.
- Meeter, M., Murre, J. M. J., & Talamini, L. M. (2002). A computational approach to memory deficits in schizophrenia. *Neurocomputing*, 44, 929-936.
- Meeter, M., Talamini, L. M., & Murre, J. M. J. (2004). Mode shifting between storage and recall based on novelty detection in oscillating hippocampal circuits. *Hippocampus*, 14, 722-741.
- Miller, R. R., & Matzel, L. D. (1988). The comparator hypothesis: A response rule for the expression of associations. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 22, pp. 51-92). New York: Academic Press.
- Moore, J., Goodell, N., & Solomon, P. (1976). Central cholinergic blockade by scopolamine and habituation, classical conditioning, and latent inhibition of the rabbit's nictitating membrane response. *Physiological Psychology*, 4(3), 395-399.
- Murnane, K., Phelps, M. P., & Malmberg, K. (1999). Context-dependent recognition memory: The ICE theory. *Journal of Experimental Psychology: General*, 128, 403-415.
- Murre, J. M. J. (1996). TraceLink: A model of amnesia and consolidation of memory. *Hippocampus*, 6, 675-684.
- Myers, C. E., Ermita, B., Harris, K., Hasselmo, M., Solomon, P., & Gluck, M. (1996). A computational model of the effects of septohippocampal disruption on classical eyeblink conditioning. *Neurobiology of Learning and Memory*, 66, 51-66.

- Myers, C. E., & Gluck, M. (1996). Cortico-hippocampal representations in simultaneous odor discrimination learning: A computational interpretation of Eichenbaum, Mathews & Cohen (1989). *Behavioral Neuroscience*, *110*(4), 685-706.
- Myers, C. E., & Gluck, M. A. (1994). Context, conditioning and hippocampal re-representation. *Behavioral Neuroscience*, *108*, 835-847.
- Myers, C. E., Gluck, M. A., & Granger, R. (1995). Dissociation of hippocampal and entorhinal function in associative learning: A computational approach. *Psychobiology*, *28*, 116-138.
- Nadel, L., & Moscovitch, M. (1997). Memory consolidation, retrograde amnesia and the hippocampal complex. *Current Opinion in Neurobiology*, *7*, 217-227.
- Nadel, L., Samsonovitch, A., Ryan, L., & Moscovitch, M. (2000). Multiple trace theory of human memory: Computational, neuroimaging and neuropsychological results. *Hippocampus*, *10*, 352-368.
- Norman, K. A., & O'Reilly, R. C. (2003). Modeling Hippocampal and Neocortical Contributions to Recognition Memory: A Complementary Learning Systems Approach. *Psychological Review*, *110*, 611-646.
- Ogden, J. A. (1993). Visual object agnosia, prosopagnosia, achromatopsia, loss of visual imagery, and autobiographical amnesia following recovery from cortical blindness: Case M.H. *neuropsychologia*, *31*, 571-589.
- Oja, E. (1982). A simplified neuron model as a principal component analyzer. *Journal of Mathematical Biology*, *15*, 267-273.
- O'Keefe, J. (1979). A review of the hippocampal place cells. *Progress in Neurobiology*, *13*, 419-439.
- O'Reilly, R. C. (1996). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation*, *8*, 895-938.
- O'Reilly, R. C., & Rudy, J. W. (2000). Computational principles of learning in the neocortex and hippocampus. *Hippocampus*, *10*, 389-397.
- 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.
- Pagnoni, G., Zink, C. F., Montague, P. R., & Berns, G. S. (2002). Activity in human ventral striatum locked to errors of reward prediction. *Nature Neuroscience*, *5*, 97-98.
- Pearce, J. M., & Bouton, M. E. (2001). Theories of associative learning in animals. *Annual Review of Psychology*, *52*, 111-139.
- Pearce, J. M., & Hall, G. (1980). A model for Pavlovian learning: Variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychological Review*, *87*, 532-552.
- Penick, S., & Solomon, R. (1991). Hippocampus, context and conditioning. *Behavioral Neuroscience*, *105*(5), 611-617.
- Peoples, L. L., Uzwiak, A. J., Gee, F., & West, M. O. (1997). Operant behavior is necessary and sufficient of rat accumbens neurons during sessions of intravenous cocaine infusions. *Brain Research*, *757*, 280-284.
- Raaijmakers, J. G. W., & Shiffrin, R. M. (1992). Models for recall and recognition. *Annual Review of Psychology*, *43*, 205-234.
- Ranganath, C., Yonelinas, A. P., Cohen, M. X., Dy, C. J., Tom, S. M., & D'Esposito, M. (2004). Dissociable correlates of recollection and familiarity within the medial temporal lobes. *Neuropsychologia*, *42*, 2-13.
- Reber, P. J., & Squire, L. R. (1998). Contrasting cortical activity associated with category memory and recognition memory. *Learning & Memory*, *5*, 420-428.
- Recanzone, G. H., Schreiner, C. E., & Merzenich, M. M. (1993). Plasticity in the frequency representation of primary auditory cortex following discrimination training in adult owl monkeys. *Journal of Neuroscience*, *13*, 87-103.

INCREMENTAL LEARNING & EPISODIC MEMORY

- Reed, J. M., & Squire, L. R. (1998). Retrograde amnesia for facts and events: Findings from four new cases. *Journal of Neuroscience*, *18*, 3943-3954.
- Rempel-Clower, N. A., Zola, S. M., Squire, L. R., & Amaral, D. G. (1996). Three cases of enduring memory impairment after bilateral damage limited to the hippocampal formation. *Journal of Neuroscience*, *16*, 5233-5255.
- Rescorla, R., & Wagner, A. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and non-reinforcement. In A. B. a. W. Prokasy (Ed.), *Classical Conditioning II: Current Research and Theory* (pp. 64-99). New York: Appleton-Century-Crofts.
- Robertson, I. H., & Murre, J. M. J. (1999). Why do damaged brains recover? A neuropsychological analysis in a connectionist framework. *Psychological Bulletin*, *125*, 544-575.
- Rokers, B., Mercado, E., Allen, M. T., Myers, C. E., & Gluck, M. A. (2002). A connectionist model of septohippocampal dynamics during conditioning: Closing the loop. *Behavioral Neuroscience*, *116*, 48-62.
- Rudy, J. W., & Sutherland, R. (1995). Configural association theory and the hippocampal formation: An appraisal and reconfiguration. *Hippocampus*, *5*, 375-398.
- Sah, P., & Clements, J. D. (1999). Photolytic manipulation of [CA2+] reveals slow kinetics of potassium channels underlying the afterhyperpolarization in hippocampal pyramidal neurons. *Journal of Neuroscience*, *19*, 3657-3664.
- Schmajuk, N., & DiCarlo, J. (1992). Stimulus configuration, classical conditioning and hippocampal function. *Psychological Review*, *99*, 268-305.
- Schmaltz, L., & Theios, J. (1972). Acquisition and extinction of a classically conditioned response in hippocampectomized rabbits (*Oryctolagus cuniculus*). *Journal of Comparative and Physiological Psychology*, *79*, 328-333.
- Schultz, W. (1998). Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, *80*, 1-27.
- Schultz, W. (2002). Getting formal with dopamine and reward. *Neuron*, *36*, 241-263.
- Scoville, W. B., & Milner, B. (1957). Loss of recent memories after bilateral hippocampal lesions. *Journal of Neurology, Neurosurgery and Psychiatry*, *20*, 11-21.
- Shohamy, D., Allen, M., & Gluck, M. (2000). Dissociating entorhinal and hippocampal involvement in latent inhibition. *Behavioral Neuroscience*, *114*(5), 867-874.
- Smith, S. M., & Vela, E. (2001). Environmental context-dependent memory: A review and meta-analysis. *Psychonomic Bulletin & Review*, *8*, 2003-2220.
- Sohal, V. S., & Hasselmo, M. E. (2000). A model for experience-dependent changes in the responses of entorhinal-temporal neurons. *Network: Computations in Neural Systems*, *11*, 169-190.
- Squire, L. R., & Alvarez, P. (1995). Retrograde amnesia and memory consolidation: a neurobiological perspective. *Current Opinion in Neurobiology*, *5*, 169-175.
- Squire, L. R., Cohen, N. J., & Zola-Morgan, M. (1984). The medial temporal lobe memory system. In H. Weingarter & E. Parker (Eds.), *Memory consolidation* (pp. 185-210). Hillsdale, NJ: Lawrence Erlbaum.
- Steinmetz, J. (1998). The localization of a simple type of learning and memory: The cerebellum and classical eyeblink conditioning. *Current Directions in Psychological Science*, *7*(3), 72-77.
- Sutherland, R., & Rudy, J. (1989). Configural association theory: The role of the hippocampal formation in learning, memory and amnesia. *Psychobiology*, *17*(2), 129-144.
- Suzuki, W. A., Miller, E. K., & Desimone, R. (1997). Object and place memory in the macaque entorhinal cortex. *J Neurophysiol*, *78*(2), 1062-1081.

- Symonds, M., & Hall, G. (1995). Perceptual learning in flavor aversion conditioning: Roles of stimulus comparison and latent inhibition of common stimulus elements. *Learning and Motivation, 26*, 203-219.
- Talamini, L. M., Meeter, M., Murre, J. M. J., Elvevåg, B., & Goldberg, T. E. (in press). Integration of parallel input streams in parahippocampal model circuits; implications for schizophrenia. *Archives of General Psychiatry*.
- Teyler, T. J., & DiScenna, P. (1986). The hippocampal memory indexing theory. *Behavioral Neuroscience, 100*, 147-154.
- Thomas, M. J., & Malenka, R. C. (2003). Synaptic plasticity in the mesolimbic dopamine system. *Philosophical Transactions of the Royal Society, London: Biological sciences, 358*, 815-819.
- Thompson, R. (1990). Neural mechanisms of classical conditioning. *Philosophical Transactions of the Royal Society, London [Biology], 329*, 161-170.
- Thompson, R., & Gluck, M. (1991). Brain substrates of basic associative learning. In H. W. a. R. Lister (Ed.), *Cognitive Neuroscience* (pp. 24-45). New York: Oxford University Press.
- Tranel, D., Damasio, H., & Damasio, A. R. (2000). Amnesia caused by herpes simplex encephalitis, infarctions in basal forebrain, and anoxia/ischemia. In F. Boller & J. Grafman (Eds.), *Handbook of Neuropsychology (2nd Ed.)*, Vol. 2. Amsterdam: Elsevier Science.
- Tsujimoto, S., & Sawaguchi, T. (2004). Properties of delay-period neuronal activity in the primate prefrontal cortex during memory- and sensory-guided saccade tasks. *European Journal of Neuroscience, 19*, 447-457.
- Vann, S. D., Brown, M. W., Erichsen, J. T., & Aggleton, J. P. (2000). Fos imaging reveals differential patterns of hippocampal and parahippocampal subfield activation in rats in response to different spatial memory tests. *The Journal of Neuroscience, 20*(7), 2711-2718.
- Vinogradova, O. S., Kitchigina, V. F., & Zenchenko, C. I. (1998). Pacemaker neurons of the forebrain medial septal area and theta rhythm of the hippocampus. *Membrane and Cell Biology, 11*(6), 715-725.
- Wagner, A. (1981). SOP: A model of automatic memory processing in animal behavior. In N. S. a. R. Miller (Ed.), *Information Processing in Animals: Memory Mechanisms* (pp. 5-47). Hillsdale, NJ: Erlbaum.
- Witter, M. P., Wouterlood, F. G., Naber, P. A., & Van Haeften, T. (2000). Anatomical organization of the parahippocampal-hippocampal network. *Annals of the New York Academy of Sciences, 911*, 1-24.
- Xiang, J. Z., & Brown, M. W. (1998). Differential neuronal encoding of novelty, familiarity, and recency in regions of the anterior temporal lobe. *Neuropharmacology, 37*, 657-676.
- Yonelinas, A. P. (2001). Components of episodic memory: The contribution of recollection and familiarity. *Philosophical Transactions of the Royal Society of London B, 356*, 1363-1374.
- 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.
- Yu, Q. X., Wang, J. J., & Chen, J. (1989). Hippocampus-cerebellar cortex-cerebellar nuclei projection in the rat: Electrophysiological and HRP studies. *Shen Li Xue Bao, 41*, 231-240.
- Zhu, X. O., McCabe, B. J., Aggleton, J. P., & Brown, M. W. (1997). Differential activation of the rat hippocampus and perirhinal cortex by novel visual stimuli and a novel environment. *Neuroscience Letters, 229*(2), 141-143.

Appendix

Representational layers

Both the cortical and the parahippocampal layers consist of 33 nodes, enough to code for 2 contexts of 15 nodes, 2 stimuli of 1 node and a US. The hippocampal layer consists of 40 nodes. Weights between the cortical and parahippocampal layers are of two kinds. Each cortical node has an immutable connection with weight 1.8 with one parahippocampal node. With all other parahippocampal nodes, they have connections subject to learning that are initialized at values taken from a uniform distribution with mean .05 and a 25% spread. Parahippocampal nodes are in turn connected to 85% of hippocampal nodes with connections subject to learning; weights on these connections are initialized at values taken from a uniform distribution with mean .3 and a 10% spread.

The parahippocampal and hippocampal layers consist of simple firing rate nodes that compute an output from linearly summated inputs. Input to the cells, g_i , consists of excitation, E_i , and inhibition, I_i , both originating from the layer below. Excitation, unique to each node, is a weighted sum of the output of the layer below; the inhibitory component, modeling undifferentiated feedforward inhibition, is a fraction of the summed output, the same for all nodes:

$$g_i = E_i - I_i = \Sigma(w_{ij}o_j) - \eta \Sigma(o_j)$$

where o_j is the output of presynaptic node j , w_{ij} the weight on its connection to node i . The inhibition fraction η is set to a value 1.2 times the average weight at the outset of the simulation (0.06 from cortex to EC, and 0.36 from EC to the hippocampus), so as to strongly inhibit nodes not part of a cued memory. A node's output is a function of its input g_i , and through adaptation of its own previous firing record. Using t as the time index:

$$o_i(t) = g_i(t) * [1 - \alpha A_i(t)] * H[g_i(t)] \quad \text{and}$$

$$\Delta A_i(t) = -\tau A_i(t-1) + \beta o_i(t-1)$$

H is the Heavyside step function that is 0 below 0 and 1 above that value, and $1 - \alpha A(t)$ is an adaptation function. The accretion parameter β of adaptation is set to 0.5, its gain parameter α to 0.9, and its decay constant τ to 0.1. This implies a slow recovery from adaptation (seconds instead of milliseconds), as is indeed present in at least some subtypes of neocortical (Fleidervish, Friedman, & Gutnick, 1996) and hippocampal (Sah & Clements, 1999) neurons. Learning is implemented using Oja's rule:

$$\Delta w_{ij} = \mu^+(o_j o'_i)(1 - w_{ij}) - \mu^-(1 - o_j) o'_i w_{ij}$$

INCREMENTAL LEARNING & EPISODIC MEMORY

This variant of Hebb's rule seems to describe LTP well (Levy, Colbert, & Desmond, 1990) and allows learning to asymptotically approach a plateau whose level determined by the strength of the input o_j , and the balance between μ^+ and μ^- . A modified output level o'_i is used in the learning rule, which is equal to $o_i - 0.4$, truncated at zero. Only strongly firing nodes are thus allowed to learn. This models data showing that the threshold for LTP is usually higher than the firing threshold.

The positive learning rate μ^+ is set at 0.01 for the connections from the cortical layer to the parahippocampal layer (except in the first familiarity simulation, where it was set at 0.1). For the connections from the parahippocampal layer to the hippocampal layer, it is set at 0.05. The negative rate μ^- was set to $0.5 * \mu^+$ in all connections.

Septal activity

Learning under influence of ACh is implemented in the connection of the parahippocampus to the hippocampus. This was not done in great detail. Learning under influence of ACh is triggered whenever fewer than three hippocampal nodes has an output of more than 0.05 (i.e., when patterns were new), or whenever prediction error in the cerebellum is higher than 0.5. In these cases, the 6 nodes with the highest inputs are allowed to strengthen the weights on their connections from the parahippocampal layers according to the following rule:

$$\Delta w_{ij} = \Psi(o_j o_i)(1 - w_{ij}) - 0.5 \Psi(1 - o_j) w_{ij}$$

Here o_i is the output of the i 'th hippocampal node, o_j the output of the j 'th parahippocampal node, and Ψ the cholinergic input (10 in case of a novel pattern, $1.5 * (\text{error} - 0.5)$ in case of septal activation through error). Weight change is bound to the interval $[-0.25, 0.3]$.

Cerebellum

The cerebellar network is the same as that in earlier work (Gluck, Myers, & Thompson, 1994), consisting of a simple perceptron trained with the Rescorla Wagner rule:

$$\Delta w_{oi} = \alpha_c (o_o - t_o) o_i$$

where $(o_o - t_o)$ is the error of the output node (its output minus its target), and o_i is the output of the i 'th input node. The target of the output node is the presence of a US on the next iteration. (1 if present, 0 if absent). There are 99 input nodes, with 33 getting input from each 1 context node (transmitted via weights of strength 1), 33 more getting input from each 1 parahippocampal node (transmitted via weights of strength 1), and the last 33 getting connections with strength 0.5 from each a random 15% of hippocampal nodes. Weights between the input layer and output node are initialized at values taken from a uniform distribution between -0.05 and 0.05 . The learning rate α_c is set at 0.05 in all simulations.

Figure 1 The cortico-hippocampal model of Gluck and Myers (1993). The model receives inputs representing conditioned stimuli (CSs), such as lights and tones, as well as contextual information. One network, representing processing which is dependent on the hippocampal-region, learns to reconstruct these inputs and also to predict arrival of the unconditioned stimulus (US), such as a corneal airpuff. As it does, the hippocampal-region network forms new stimulus representations in its internal layer that compress redundant information and differentiate predictive information. A second network, assumed to represent long-term memory sites in cerebral and cerebellar cortices, adopts the internal representation provided by the hippocampal-region network, and then to map from this to an output that represents the strength or probability of a conditioned response (CR), such as a protective eyeblink.

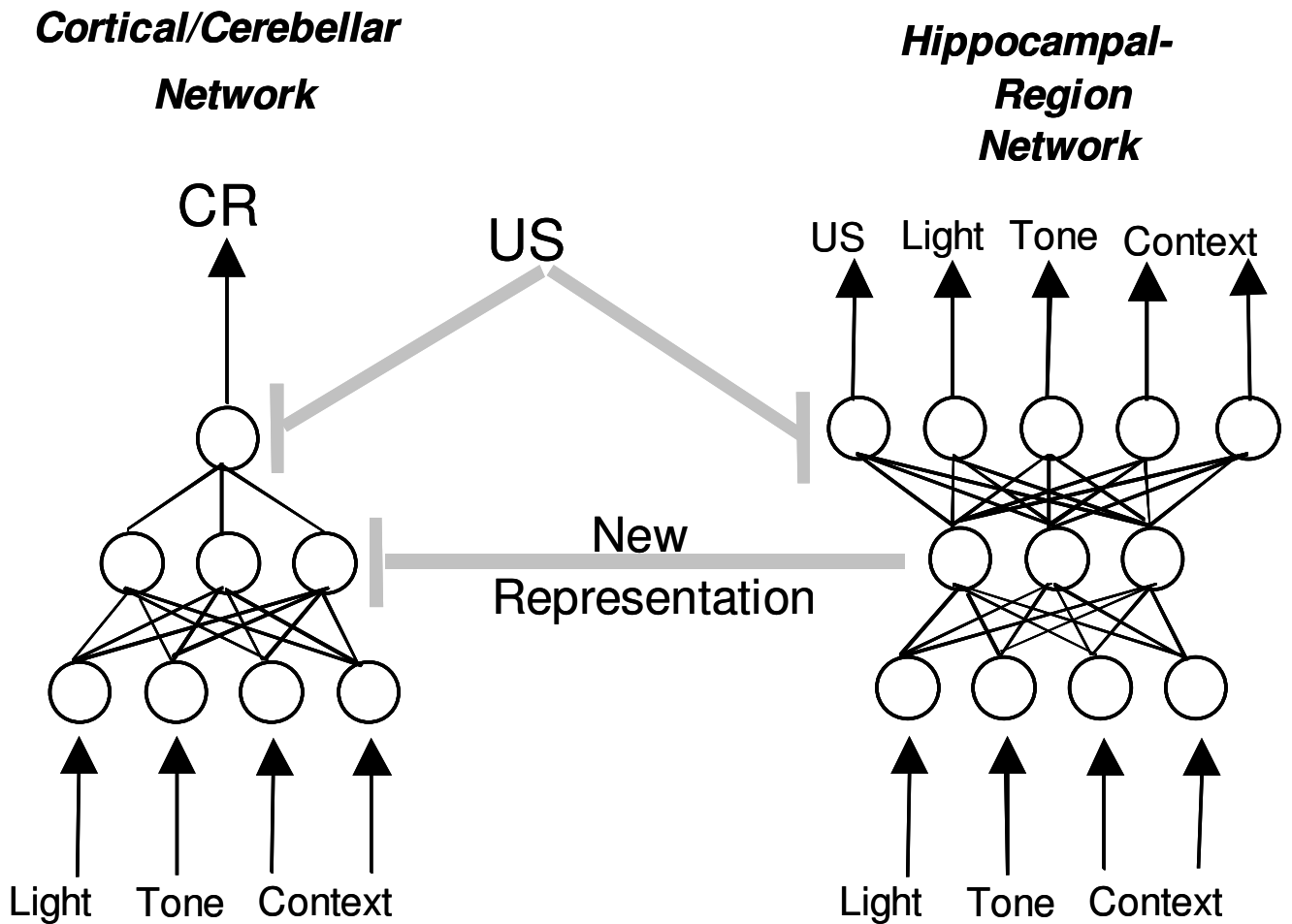


Figure 2: Brain regions that play a part in incremental learning. See main text for explanation.

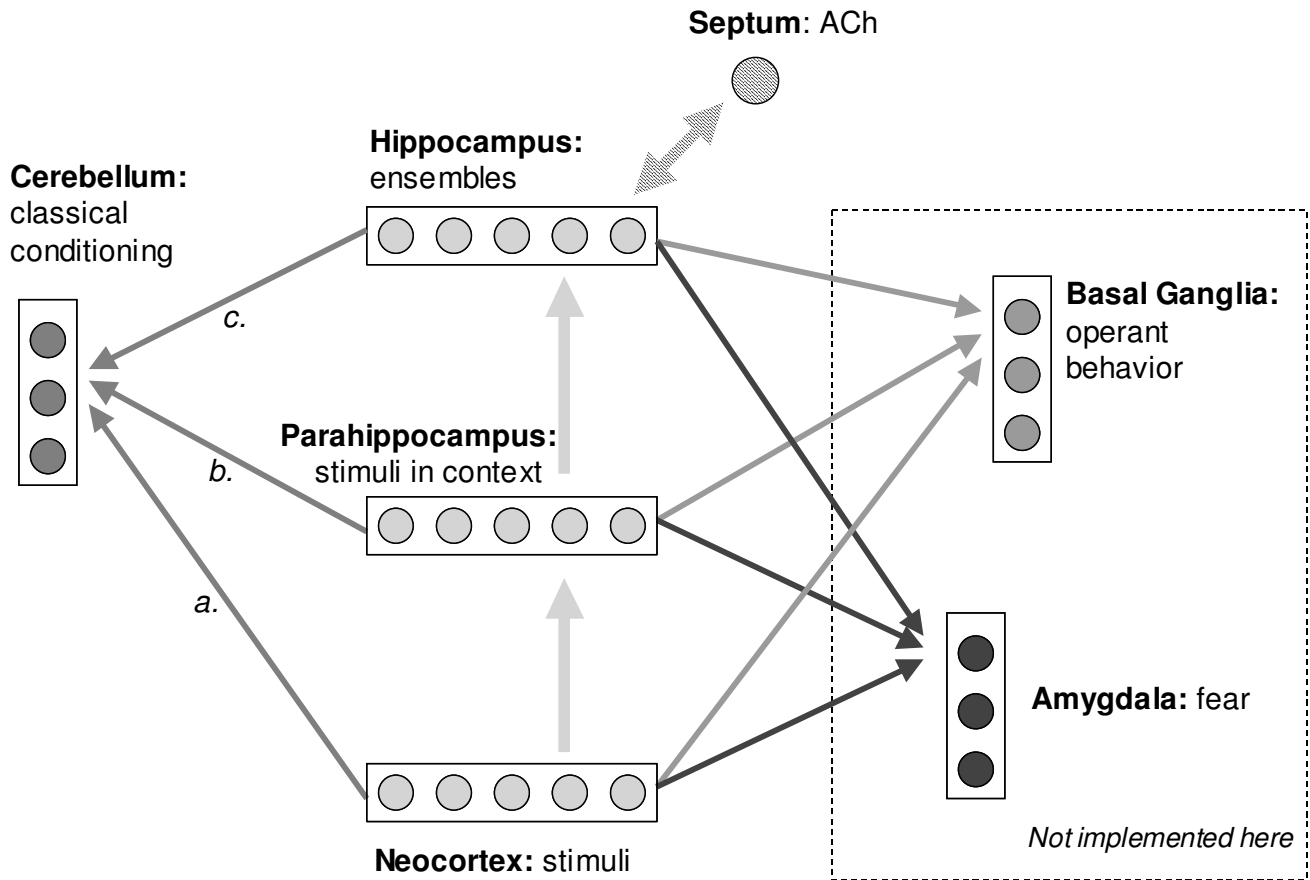


Figure 3 Neuronal responses of a familiarity neuron, in this case located in area TE, to a visual stimulus presented either for the first time (left panel), or repeated in a different session (first presentation in that session, right panel). Reprinted with permission from Xiang and Brown (1998).

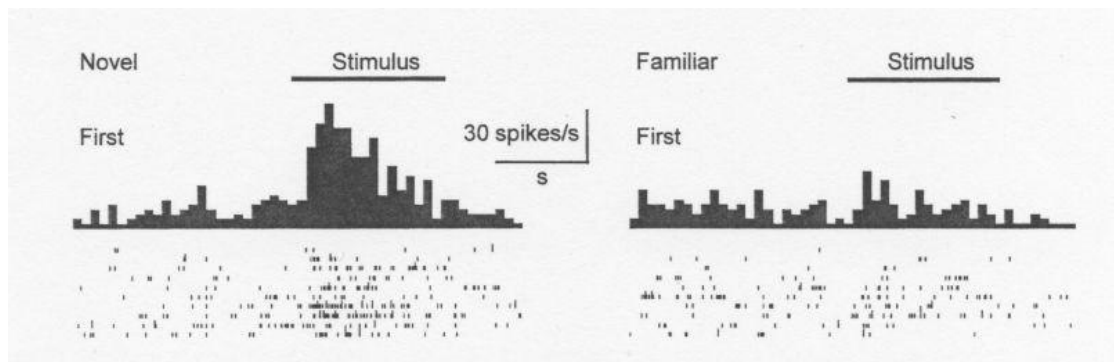


Figure 4: Theory of the familiarity effect. Some cortical nodes code for features that are more or less permanent in the current environment (labeled 'context'), others code for phasic cues (labeled 'stimulus'). Each cortical node has a strong connection to one parahippocampal node. In addition, they send weak connections to other parahippocampal nodes (connections to only one parahippocampal node are drawn). (a). When a phasic cue is presented together with a set of more permanent context features, cortical nodes coding for the context features are active synchronously with parahippocampal nodes activated by the phasic cue. This will allow LTP to occur in the connections from the cortical nodes to the parahippocampal nodes. (b). The strengthened connections now allow the permanent context features to weakly activate the parahippocampal node on their own. (c). The resulting adaptation in the parahippocampal node will make it less responsive to its preferred input, the phasic cue, which produces the familiarity effect.

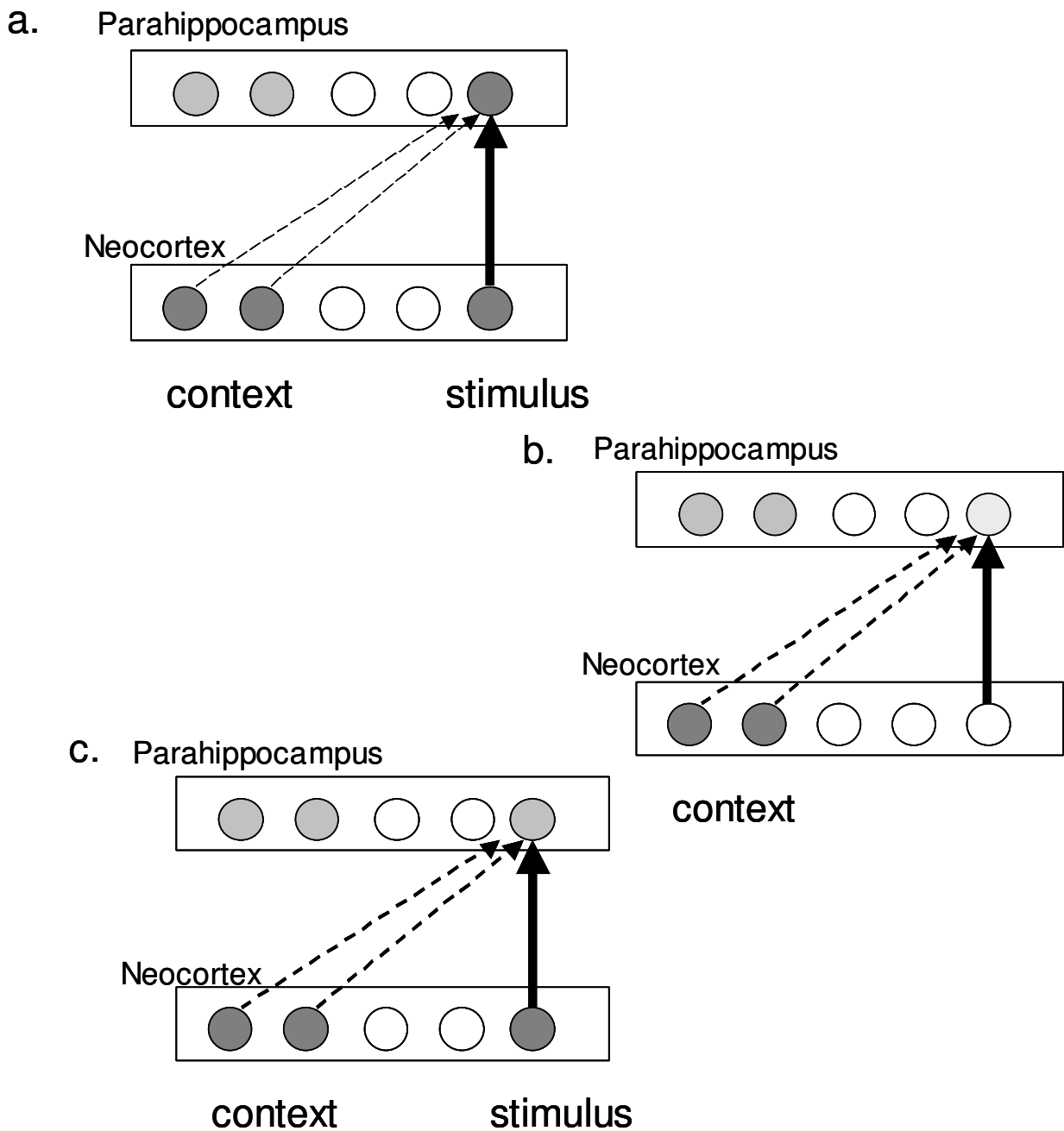


Figure 5 Recollection and familiarity in the Norman and O'Reilly (2003) model. Left panel shows how the familiarity (famil.) and recollection (recoll.) signals to a probe decrease with decreasing overlap between the probe and previously stored inputs. Right panel: strength of both signals to studied items (targets), similar lures (sim. lures) that overlap in 80% of features with targets, and dissimilar lures (dissim. lures) that overlap in 20% of features with targets. Left panel plotted with data from Norman and O'Reilly (2003).

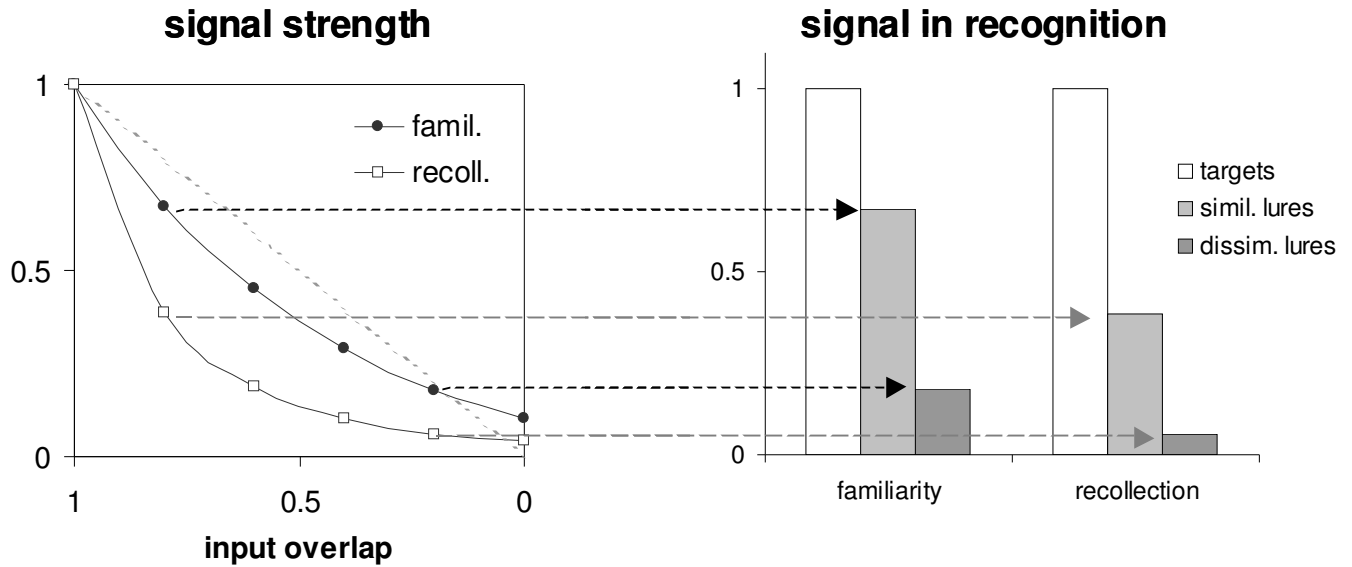
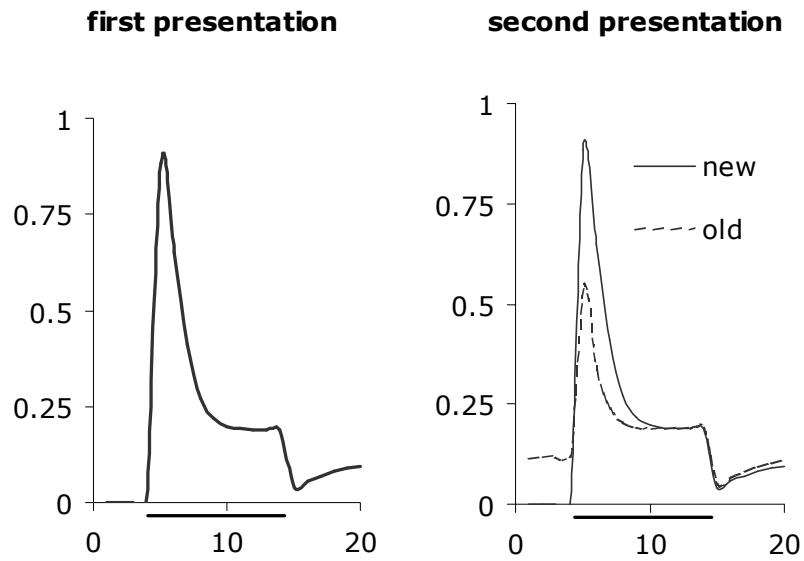


Figure 6 Parahippocampal coding of stimulus familiarity. (a). Response of the parahippocampal node to its preferred stimulus, with rapid learning from the cortical layer to the parahippocampal layer ($\mu=0.1$). To the left the node's response to a stimulus presented for the first time in a context (black bar). To the right a comparison of stimuli presented after this stimulus in the same context. If a novel stimulus is presented ('new'), the response of the parahippocampal node coding for this stimulus is of the same magnitude. If the first stimulus is repeated ('old'), the response is attenuated. (b). Response of a parahippocampal node to its preferred repeated 100 times in the same context, with slow learning from the cortical layer to the parahippocampal layer ($\mu=0.01$). The response slowly habituates with repeated presentations (x-axis= number of repetitions).

a.



b.

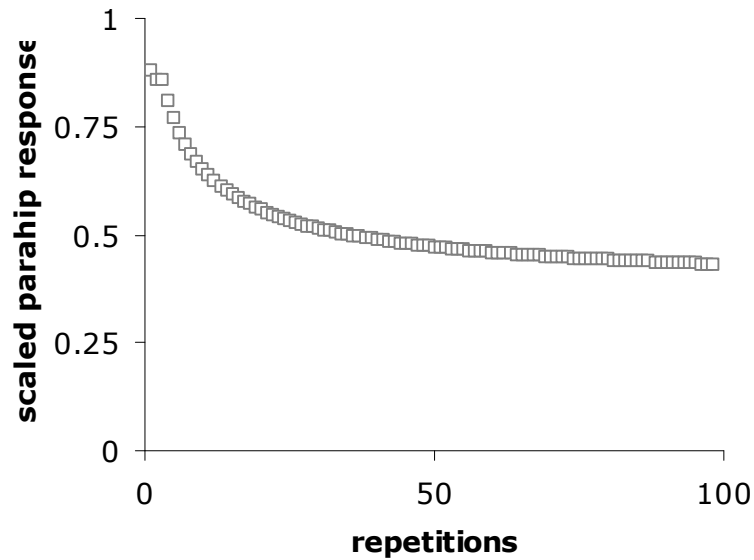
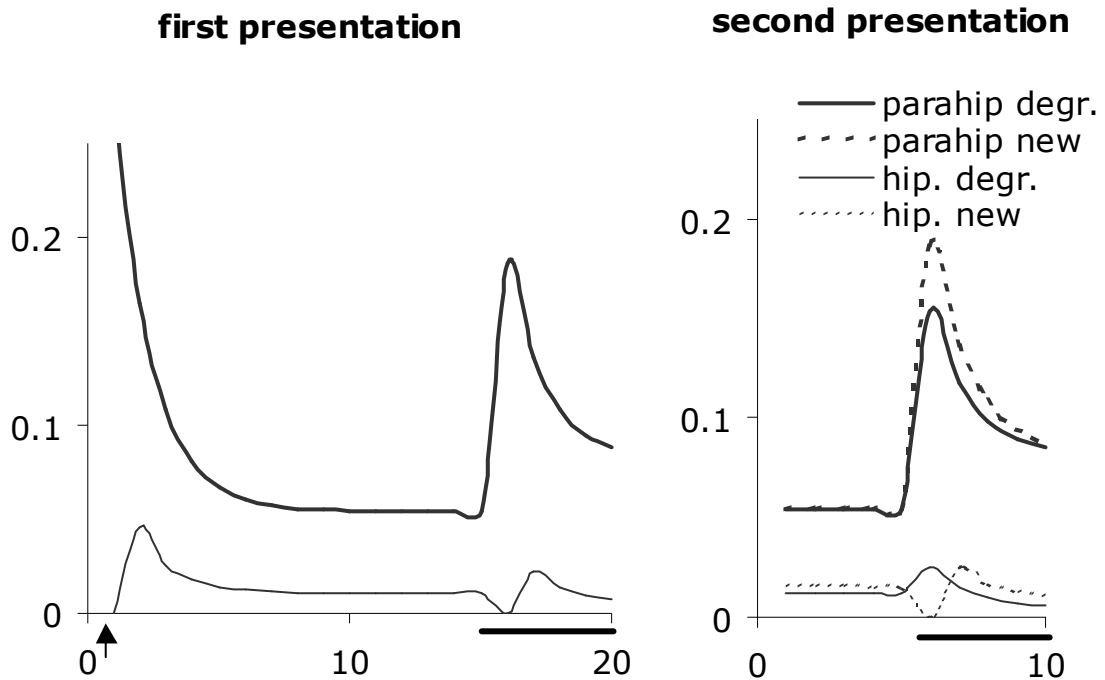


Figure 7 (a). Response of the parahippocampal and hippocampal layer (average response of all nodes in each layer) to novel and old patterns. In the left panel the first presentation of a pattern is at time step 16 (black bar). In the second panel the responses to either a novel pattern, or a degraded version of the first pattern (black bar). The degraded old pattern immediately elicits activity. Novel patterns cause a pause in firing in the hippocampus in the first time step they are presented. ACh-induced learning subsequently creates a pattern, which is active in later time steps. The same is true for the first presentation of the context, at time step 1 (black arrow) (b). Correlation of hippocampal activity elicited by the first pattern with that elicited by either the degraded version of the first, or the novel pattern. The response to the degraded pattern resembles that to the first pattern, while the novel pattern generates an entirely new pattern of hippocampal activity, negatively correlated with the first pattern.

a.



b.

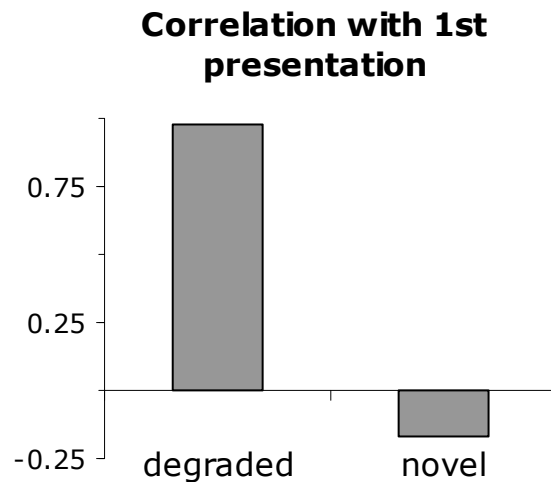
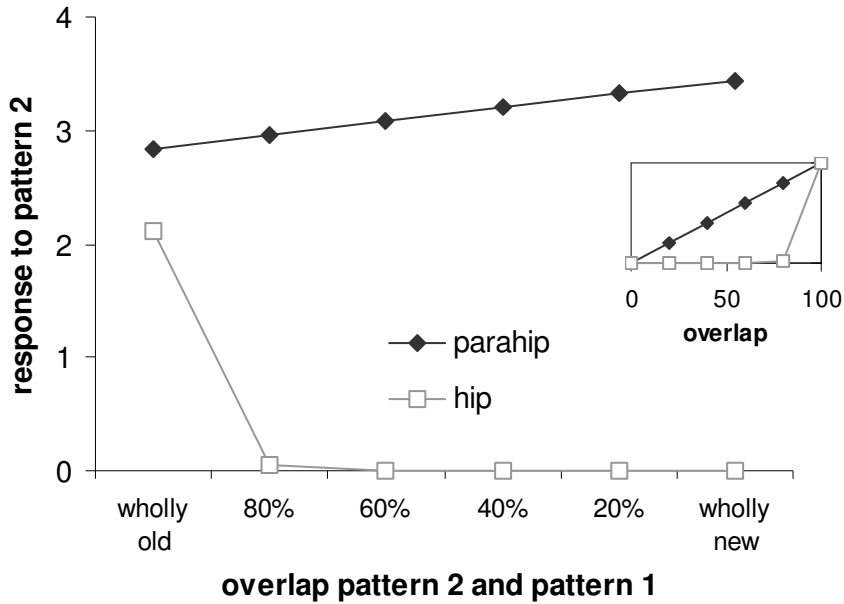


Figure 8 (a). Responses of the parahippocampal and hippocampal layer to a second pattern with varying overlap with an already stored pattern. Simulation structure as in Figure 7. Plotted is the average node output in each layer on the first time step that the second pattern is presented, divided by the average node output to context alone. Inset of (a): layer activities replotted as familiarity and recollection signals, as in Figure 5. (b). Strength of both signals to studied items (targets), similar lures (sim. lures) that overlap in 80% of features with targets, and dissimilar lures (dissim. lures) that overlap

a.



in 20% of features with targets.

b.

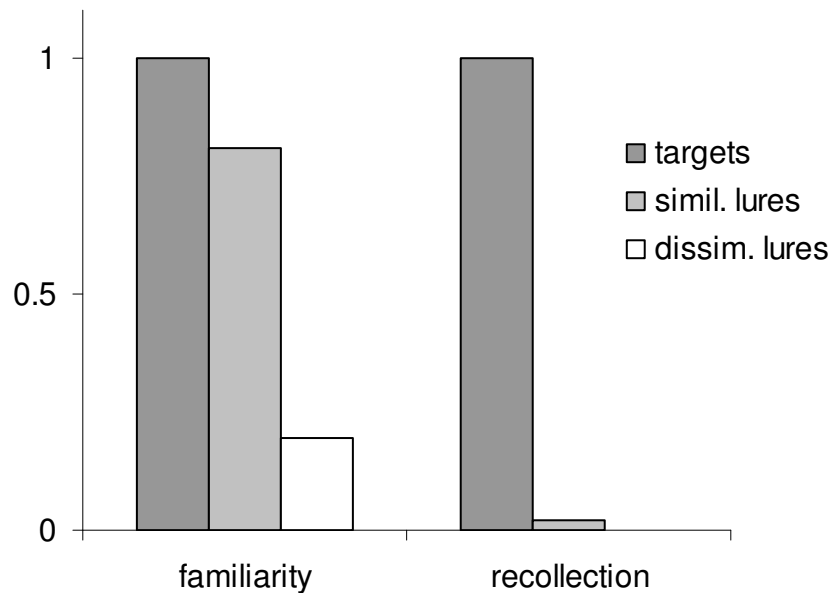


Figure 9 Acquisition of cerebellar response over 100 trials in which a CS is combined with a US. Shown is both the total cerebellar response (thick black line), and the contributions of the three sources of cerebellar inputs (hippocampal layer, parahippocampal layer and cortical layer) to this response. Right panel shows the first 10 trials, highlighting that the hippocampal contribution only starts after a few trials.

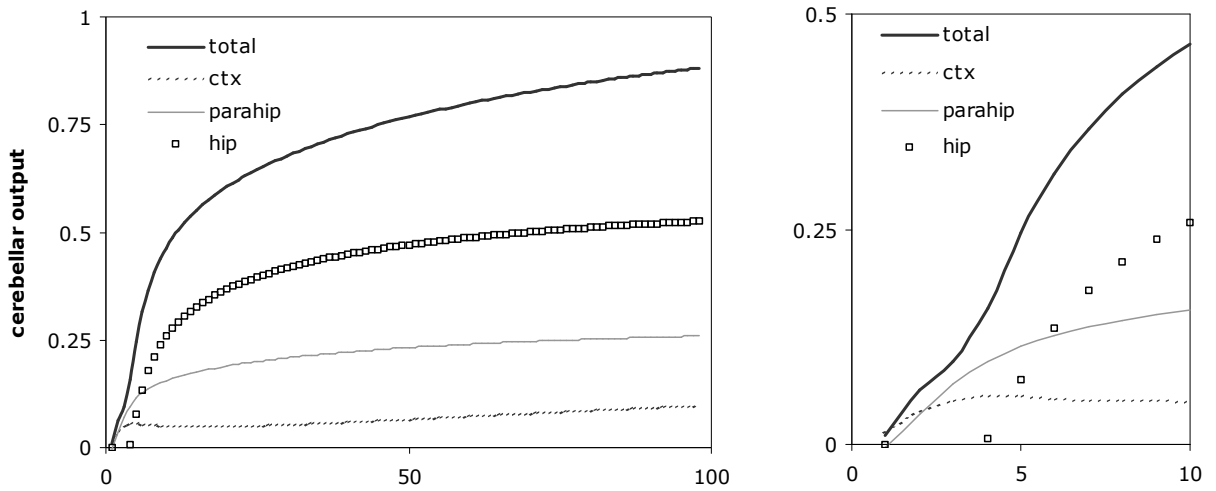


Figure 10 Cerebellar response to context alone, measured on the iteration before presentation of a stimulus. Although the total response is stable at 0, the different inputs acquire different values: hippocampal inputs provide a drive towards a CR, requiring the other inputs to actively inhibit a CR.

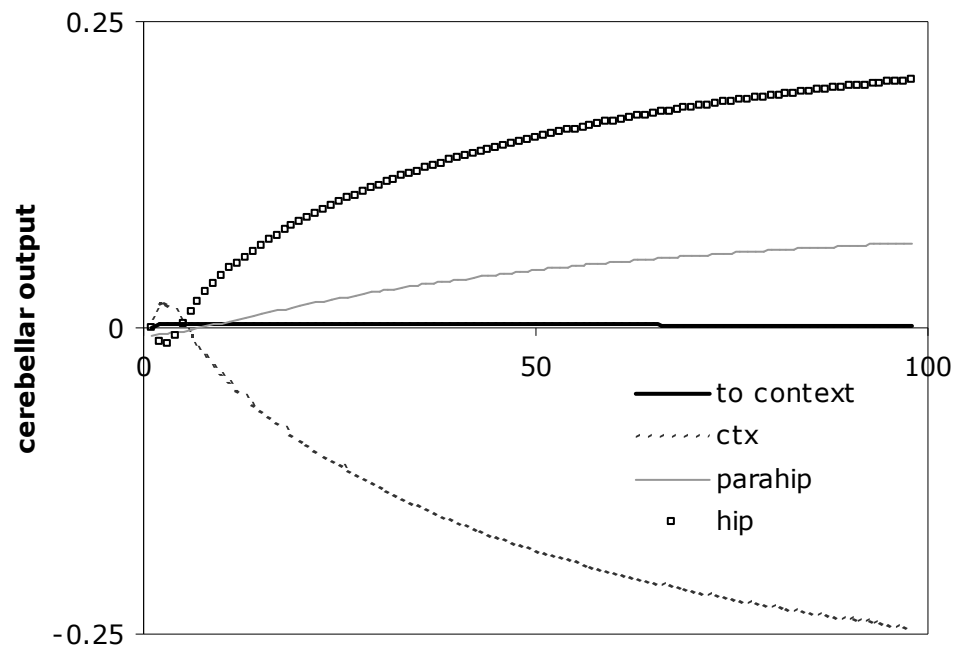
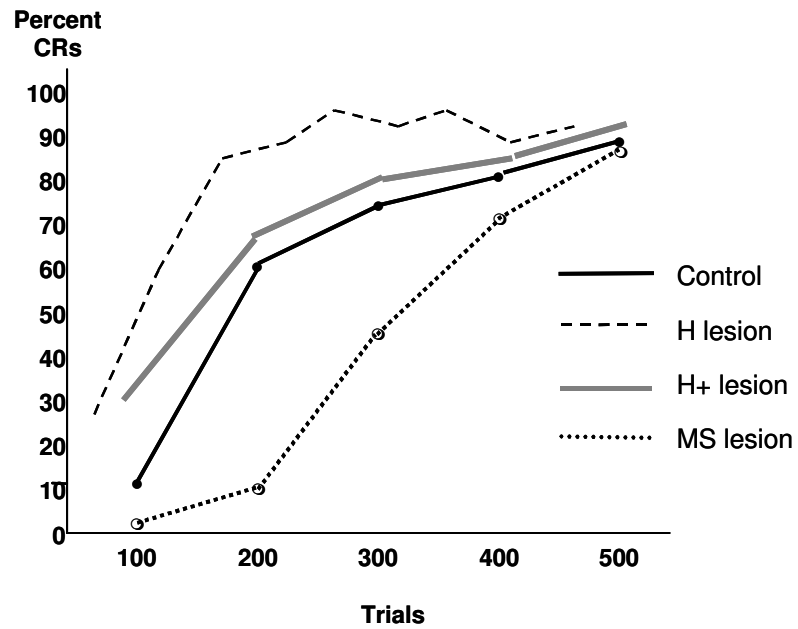


Figure 11 (a). CS-US learning in rabbit eye blink classical conditioning: only medial septal lesions (MS) reliably slow acquisition (control & MS data taken from Allen et al., 2002; HR-lesion from Allen, Chelius & Gluck, 1998; selective H lesion from Allen, Padilla, Myers & Gluck 2002; all studies used same stimulus parameters and procedure). (b) Results of simulations with classical conditioning under different lesions. As in the data, only medial septal lesions slow acquisition.

a.



b.

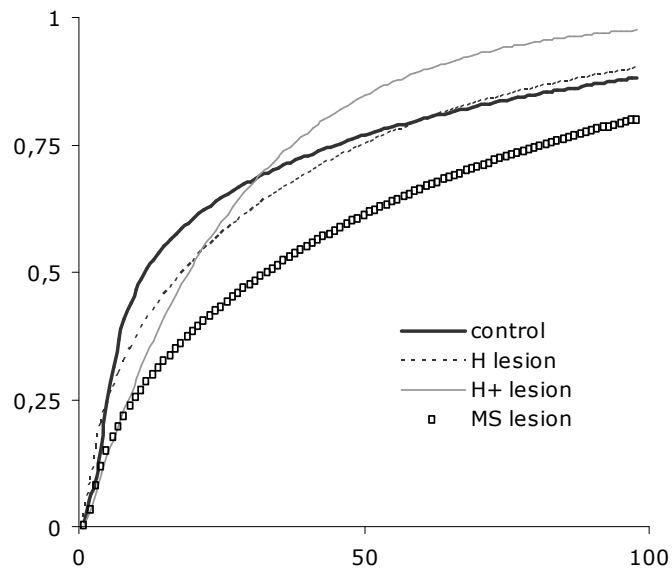
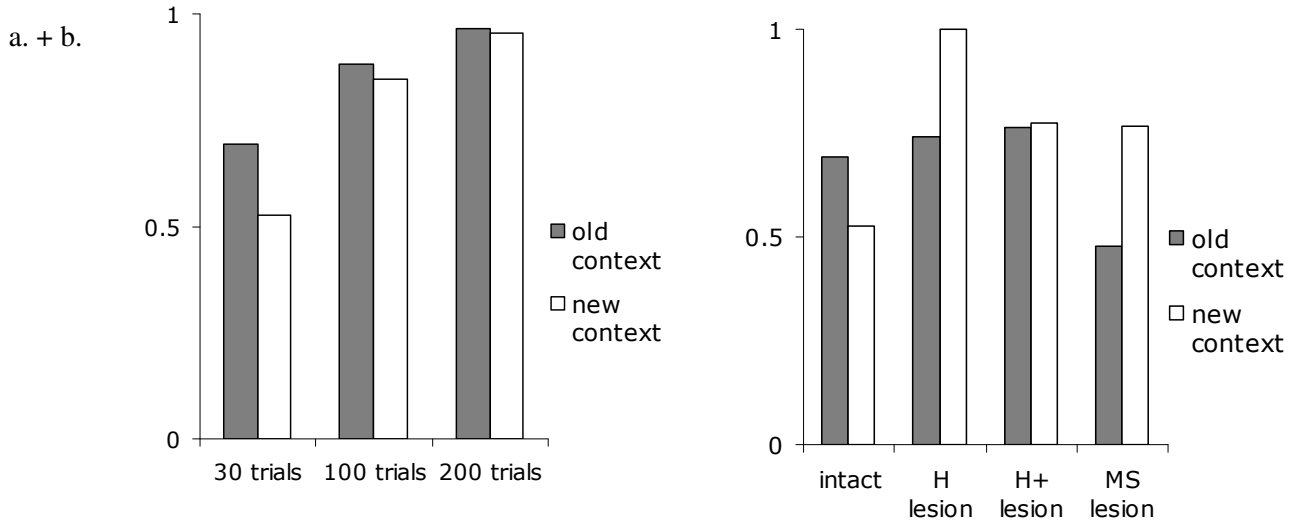


Figure 12 Effects of context change on cerebellar output. (a). In the intact model, output is smaller after a context switch. This effect is large early in training, and disappears with more trials. (b). Lesions including either the hippocampus or the medial septum abolish the context change effect. (c). In rabbit eye blink conditioning, control rabbits show decrement in learned CR when CS is presented in new context. This decrement is abolished following ablation of dorsal hippocampus (data from Penick & Solomon, 1991).



c.

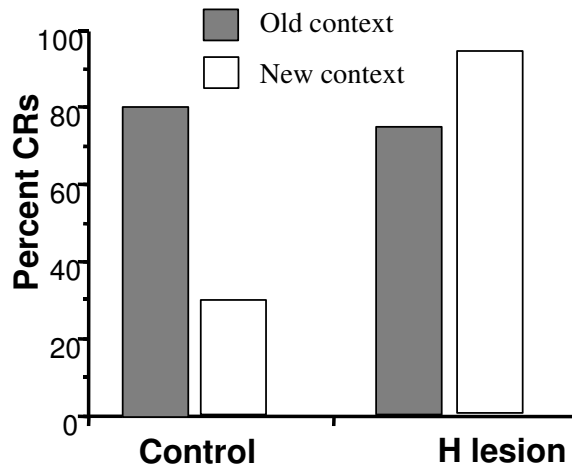


Figure 13 (a) Blocking in the intact model and all lesion conditions, as measured as the decrease in cerebellar response to B alone as compared to the response to AB. (b). Blocking in rabbit eye blink conditioning is intact in animals with selective hippocampal lesion, as measured by the decrease in responding to B in a blocking condition and in a control condition in which only B has been paired with the US (data from Allen et al. 2002) (c). The novel cue effect: the decrement in cerebellar response on the first trial that AB is presented, as compared to the last trial on which A alone is presented. (d). Novel cue effect in rabbit eye blink conditioning: decrement in first AB+ block; for both sham-lesioned controls and rabbits with selective hippocampal lesions (data from Allen, Padilla, Myers & Gluck, 2002).

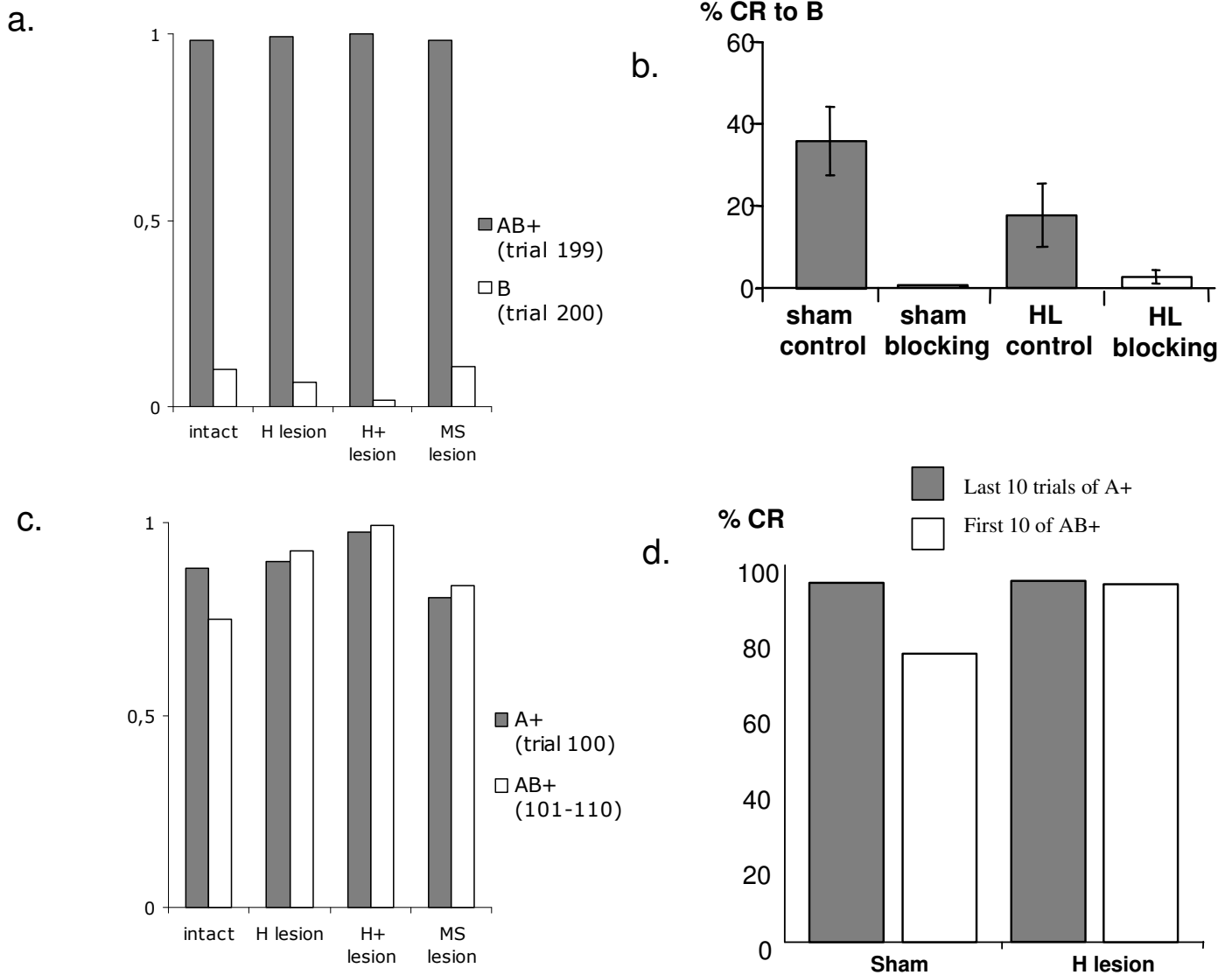


Figure 14 Acquisition of a cerebellar response in a control condition with extended context exposure, in a latent inhibition condition in which CS-US contingency is preceded by a pre-exposure trials to the CS, in a condition in which the context is changed in between pre-exposure and learning (latent inh. ctxt), and in a learned irrelevance condition in which both the CS and US are presented before learning, but in an uncorrelated fashion.

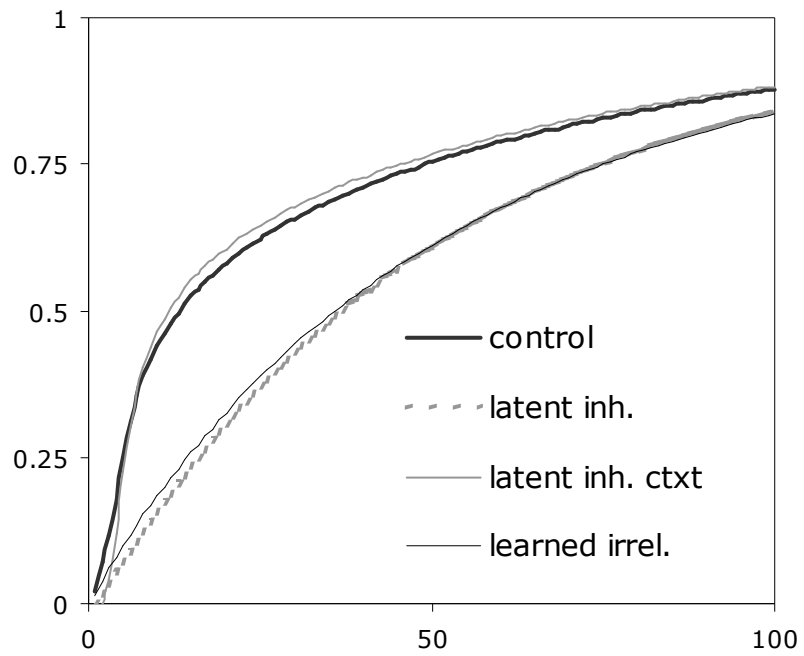


Figure 15 (a). Cerebellar output after 200 trials in a control condition, latent inhibition condition, and latent inhibition with context change condition for the intact model and all lesion models. (b). Latent inhibition in rabbit eye blink conditioning: EC lesions but not selective hippocampal lesions disrupt the effect (data from Shohamy et al., 2000). (c). Systemically administered scopolamine during CS/US exposure, which among other effects disrupts septohippocampal projections, does not disrupt learned irrelevance in rabbit eye blink conditioning (data from Moore et al., 1976).

