MEMORY MODIFICATION IN THE BRAIN: COMPUTATIONAL AND EXPERIMENTAL INVESTIGATIONS

Samuel Joseph Gershman

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of
Psychology
Advisers: Kenneth A. Norman and Yael Niv

January 2013

Abstract

I explore how and when memory traces are modified by new experience. Using a variety of paradigms, species and analytical tools, I argue that memories reflect inferences about the structure of the world. In particular, memories reflect the assignment of events to latent (hidden) causes. A new event modifies an existing memory trace if it is probable that the event was caused by the same latent cause as that represented by the old trace; otherwise, a new memory trace is formed. I show that probabilistic inference over latent causes, or *structure learning*, provides a parsimonious explanation of many phenomena in human and animal learning, and may guide us towards developing new treatments for pathological memories like trauma and addiction.

I first introduce a latent cause framework for modeling classical conditioning, based on ideas from modern Bayesian nonparametric statistics. Evidence suggests that an ostensibly extinguished memory can return under a variety of circumstances. The latent cause theory proposes that extinction training increases the probability that a new latent cause is active, thereby leading to the formation of two memories (one for acquisition, one for extinction). This theoretical explanation can also account for several other behavioral phenomena, as well as developmental trajectories and damage to the hippocampus. I argue that immature or hippocampally-damaged animals are impaired at expanding their repertoire of latent causes. I then develop a variant of the latent cause framework designed to explain the phenomenon of memory reconsolidation: retrieving a memory appears to render it temporarily labile. I show that the major phenomena of reconsolidation can be explained in terms of this framework, and I present new experimental data testing some of the theory's predictions.

Motivated by this computational framework, I explore in several experiments the factors governing latent causal inferences by rats and humans. Taken together, these experimental and theoretical results support the idea that memory modification can be understood as a process of structure learning.

Acknowledgements

This dissertation was made possible by the intensive collaboration of my advisors, Ken Norman and Yael Niv. They provided just the right balance of freedom and guidance for me to thrive at Princeton. My thinking has also been shaped by two "unofficial advisers," Dave Blei and Nathaniel Daw, who introduced me to many of the ideas discussed in this dissertation. Another important force in my personal and intellectual development has been Josh Tenenbaum, who inspired me and a generation of cognitive scientists to explore Bayesian models.

Before I came to grad school, I was the beneficiary of mentoring by several graduate students who expended an unreasonable amount of time teaching me skills and disabusing me of endless crazy notions—in particular, Joel Voss, Chris Summerfield, and Hedy Kober. If it weren't for them, I probably wouldn't be a neuroscientist today. I will always have time for inquisitive undergrads, my way of paying back part of that debt.

I am grateful to many others for collaboration, support, discussion and advice: John Myles White, Michael Todd, Aaron Courville, Peter Dayan, Anna Schapiro, Francisco Pereira, Elliot Ludvig, Richard Socher, Bob Wilson, Carlos Diuk, Ed Vul, Per Sederberg, Jon Cohen, Matt Botvinick, Marie Monfils, Nick Turk-Browne. The rat experiments described in this thesis would not have been possible without the stupendous efforts of Carolyn Jones. Finally, I cannot forget my parents, grandparents, brother and above all my wife Anna, whose encouragement animates these pages.

Contents

"There is no such thing as forgetting possible to the mind; a thousand accidents may, and will interpose a veil between our present consciousness and the secret inscriptions on the mind; accidents of the same sort will also rend away this veil; but alike, whether veiled or unveiled, the inscription remains for ever; just as the stars seem to withdraw before the common light of day, whereas, in fact, we all know that it is the light which is drawn over them as a veil – and that they are waiting to be revealed when the obscuring daylight shall have withdrawn."

—Thomas de Quincey, *Confessions of an English Opium Eater* (1822)

Chapter 1

Introduction: memory lost and found

Memories have the appearance of fragility: Forgetting is a common, almost daily, experience for most people. It is somewhat counter-intuitive, then, that many psychologists view memory traces as indelible records of experience, each stored without the corruption of other traces (Raaijmakers and Shiffrin, 1992). Supporting this view is evidence that retrieval interference is the primary determinant of forgetting (Crowder, 1976); Memories which appeared to be lost can be recovered given the right retrieval cues. In its strongest form, this retrieval view of forgetting suggests that "forgetting" as we commonly understand it (i.e., erasure of the original memory trace) is simply not possible.

If correct, the retrieval view has profound clinical implications. At some point in their lives, approximately 6.8% of persons in the United States develop post-traumatic stress disorder (PTSD) in response to a traumatic memory, characterized by insuppressible intrusions of the traumatic memory and a host of physiological distresses (Yehuda and LeDoux, 2007). An even larger proportion of the population meets the diagnostic criteria for drug addiction (Koob and Volkow, 2009). Since a key property

of PTSD and addiction is the formation of a maladaptive memory that resists erasure or modification, they can be thought of as "disorders of pathological memory." The retrieval view suggests that, once formed, a pathological memory can always potentially return under the right retrieval conditions. Thus, the best that clinicians can hope for is a temporary abeyance of the symptoms rather than a permanent cure.

Should we abandon hope? Not necessarily. In principle, one can never rule out a retrieval-based theory, since if no memory recovery is observed, one could argue that the necessary retrieval cues are not available. This theoretical slack derives from the fact that we do not, in general, know the necessary and sufficient retrieval conditions for memory recovery. Nonetheless, there are (at least in the Pavlovian conditioning literature) several widely accepted experimental measures of memory recovery, and I shall take the pragmatic viewpoint that memories which do not recover according to these measures are effectively erased. Thus, my goal will be to elucidate the conditions under which memories are modified or erased according to these conventional measures.

This thesis presents a theoretical framework for understanding memory modification, erasure, and recovery. According to Marr's (1982) taxonomy, the framework is situated primarily at the "computational level of analysis"—it formalizes the information processing task faced by the memory system, and derives a rational solution to this task (Anderson, 1990). As explained in more detail below, I pose the information processing task as inductive reasoning in a probabilistic generative model of the environment, for which the rational solution is Bayesian inference. I present the results of behavioral and brain imaging experiments to support various aspects of this theory. To demonstrate the generality of the theory, these experiments cut across different species, tasks and stimuli. I emphasize that this theory is truly a *framework*—its details vary to accommodate the variety of domains to which it is applied, but at its core is a set of computational ideas that are postulated to hold across domains.

1.1 Empirical background

In this section I review several veins of empirical data motivating my theoretical framework. I start with simple associative learning processes in animals, and then discuss perceptual and verbal learning in humans. I delve into these phenomena in greater depth in later chapters.

1.1.1 Learning and unlearning in Pavlovian conditioning

Pavlovian conditioning represents perhaps the simplest experimental paradigm for studying learning processes. In a canonical design, a motivationally neutral cue (the conditional stimulus, or CS) is first paired with an intrinsically aversive or appetitive outcome (the unconditional stimulus, or US); this is referred to as the *acquisition* (or *training*) phase. The animal typically acquires a conditioned response to the CS. In the *extinction* phase, the CS is presented without the US, resulting in a decrement of conditioned responding. Finally, a *test* phase follows typically one or two days later, in which the CS is again presented without the US.

The simplest assumption one could make about the learning processes underlying Pavlovian conditioning is that acquisition and extinction are complementary associative processes: An association between the CS and US is learned in the acquisition phase, and then unlearned in the extinction phase. According to this account, there is a single memory trace (storing the association), and this trace is oppositely modified by acquisition and extinction. This leads to the hypothesis that following extinction the original trace is irretrievably lost.

Bouton (2004) reviewed several lines of evidence suggesting that this hypothesis is incorrect:

- *Renewal.* If acquisition and extinction are performed in different contexts, then testing in the acquisition context or in a novel context results in elevated con-

ditioned responding compared to testing in the extinction context (Bouton and Bolles, 1979a; Bouton and King, 1983).

- *Reinstatement.* Reexposure to the US alone prior to test increases conditioned responding to the CS at test, as long as the CS is tested in the same context as US reexposure (Pavlov, 1927; Rescorla and Heth, 1975; Bouton and Bolles, 1979b).

- *Rapid reacquisition.* Introducing a second acquisition phase following extinction results in more rapid reacquisition of the conditioned response compared to initial acquisition (Napier et al., 1992; Ricker and Bouton, 1996).

- *Spontaneous recovery.* The mere passage of time between extinction and test is sufficient to increase conditioned responding (Pavlov, 1927; Rescorla, 2004).

If extinction isn't unlearning, what is it? Researchers are largely in agreement that extinction involves learning of something new, but what exactly is learned is open to debate. In Chapter 3, I present a computational model of the learning processes during extinction that clarifies some of these issues. The model is motivated by a Bayesian treatment of Pavlovian conditioning, using a statistical model that decides rationally when to modify old memory traces and when to create new memory traces. I discuss the Bayesian approach further below.

1.1.2 Reconsolidation and the transience of experimental amnesia

In the past decade, neuroscientists have begun a renewed program of research on memory modification, using a combination of conditioning paradigms first developed in the 1970s and modern pharmacological and neuroimaging techniques. The maelstrom of new activity has centered on the concept of *reconsolidation* (Spear, 1973).

In order to understand reconsolidation, we must first understand *consolidation*, the apparent time-dependent stabilization of memory traces (Muller and Pilzecker, 1900). The key finding motivating the concept of consolidation is the temporal gradient of retrograde amnesia (RA): new memories tend to be more susceptible to disruption than old memories (see Wixted, 2004, for a review). This gradient is seen both in experimental amnesia (e.g., induced by electroconvulsive shock; Quartermain et al., 1965; Kopp et al., 1966) and amnesia resulting from insult to the medial temporal lobes (Brown, 2002), although this assertion has not gone undisputed (Nadel et al., 2007). There is also evidence for a temporal gradient of RA for emotional memories in the amygdala (Schafe and LeDoux, 2000). The "standard model of consolidation" explains these findings by postulating that new memories exist in a temporarily labile state in the hippocampus until they are gradually transferred into a stable neocortical representation (Squire and Alvarez, 1995).

The standard model of consolidation was challenged by findings that ostensibly stable memories could be rendered labile by an appropriately timed "reminder" treatment (Lewis et al., 1968; Misanin et al., 1968; Mactutus et al., 1979). For example, administering electroconvulsive shock within a short time window after a single unreinforced CS presentation resulted in RA for the (putatively consolidated) CS-US association. Such reminder treatments not only made memories susceptible to interference by amnestic agents, but also allowed memories to be enhanced, for example by stimulation of the reticular formation (Devietti et al., 1977). These findings indicated that the temporal gradient of RA is at least partially determined by the activation state of a memory. When reactivated by a reminder treatment, memories must undergo a phase of reconsolidation to achieve stability.

After a flurry of experimental activity in the 1970s, this idea smoldered for several decades until Nader et al. (2000) showed, using a fear conditioning paradigm, that injection of the protein synthesis inhibitor (PSI) anisomycin into the basolateral

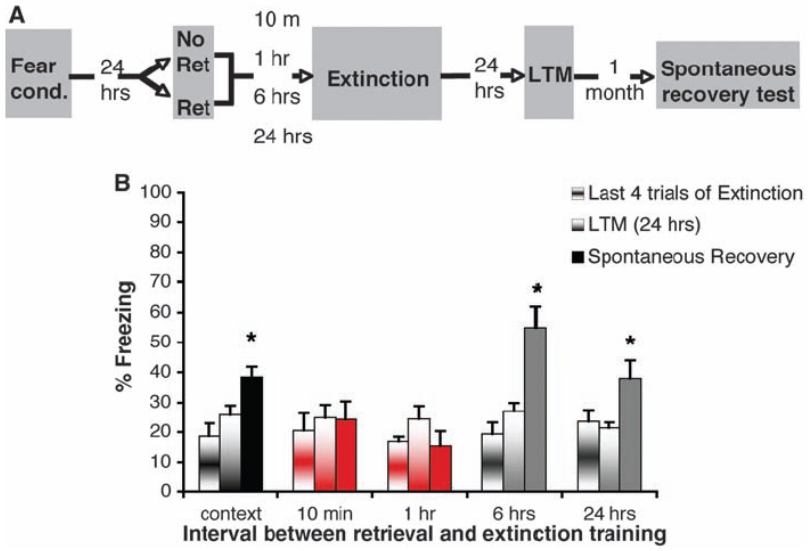

Figure 1.1: **Spontaneous recovery in the Monfils-Schiller paradigm**. Finite lability window to prevent return of fear via post-retrieval extinction. (*A*) Rats were fear-conditioned (Fear Cond) with three tone-shock pairings. After 24 hours, they were exposed either to an isolated cue retrieval trial (Ret) or context only (No Ret) followed by extinction training. The time interval between the retrieval trial (or context exposure) and the extinction was either within (10 min or 1 hour) or outside (6 hours or 24 hours) the reconsolidation window. Twenty-four hours after extinction, all groups were tested for LTM, and 1 month later for spontaneous recovery. The gray shading represents context A. (*B*) All groups were equivalent for the last four trials of extinction and at the 24-hour LTM test. One month later, the Ret groups with an interval outside the reconsolidation window (gray), as well as the No Ret group (black), showed increased freezing (spontaneous recovery) relative to the 24-hour LTM test; however, the groups with an interval within the lability window (red) did not. Error bars represent standard error. Asterisk denotes significance at the 0.05 level. Figure reproduced from Monfils et al. (2009).

amygdala following re-exposure to the training cues caused RA for the earlier fear memory. Thus, reactivated memories require new protein synthesis to reconsolidate into a stable state. This finding ushered in a new era of reconsolidation studies using pharmacological treatments (see Nader and Hardt, 2009, for a recent review). Some of the most compelling evidence for the proposition that reconsolidation induces memory modification (rather than memory formation) comes from subsequent work by Duvarci and Nader (2004) showing that several signatures of new learning during extinction reviewed above (spontaneous recovery, reinstatement, renewal) are absent following post-reactivation protein synthesis inhibition (but see Riccio et al., 2006, for discussion of evidence in favor of a retrieval deficit interpretation[1]).

1.1.3 The dynamics of memory updating

Reminders are not the only manipulations that can influence memory updating. Recent research in human psychophysics has shown that gradually morphing visual stimuli causes the stimuli to be confused in memory (Preminger et al., 2007, 2009; Wallis and Bülthoff, 2001). For example, Wallis and Bülthoff (2001) presented subjects with a rotating face that gradually morphed into a different face. Compared to a condition in which the morphs were presented in a mixed (scrambled) order, participants in the gradual morph condition were more prone to perceive the different faces as belonging to the same person as the original face.

Intriguingly, a similar phenomenon can be observed physiologically in the hippocampus. The classic work of O'Keefe and Nadel (1978) demonstrated the existence of hippocampal neurons that respond selectively when an animal occupies a partic-

[1]Re-exposing the animal to the amnestic agent can in some cases produce reactivation of the supposedly erased memory (Hinderliter et al., 1975; Briggs and Riccio, 2007; Bradley and Galal, 1988). In other words, using the amnestic agent as a retrieval cue can alleviate RA! Such findings are quite counter-intuitive if one accepts that amnestic agents degrade reactivated memories, but harmonize with the idea that learning and retrieval are "state-dependent" and that the amnestic agents function as part of the animal's internal state. Arguments against this interpretation of reconsolidation have been reviewed by Nader and Hardt (2009).

ular spatial location—*place cells*. When the animal's environment changes abruptly (as in a context switch), the place cells completely change their receptive fields, a phenomenon known as "global remapping" (Colgin et al., 2008). In contrast, smaller changes in the environment modulate the firing rates of place cells without changing their receptive fields, a phenomenon known as "rate remapping." Experiments with morphing spatial environments have reported both kinds of remapping (Wills et al., 2005; Leutgeb et al., 2005). As I discuss further in Chapter 6, which kind of remapping occurs depends on whether the morphs are gradually interpolated (in which case rate remapping occurs) or interpolated in a random order (in which case global remapping occurs).

The idea that change (or novelty) detection influences memory formation is an old one. For example, von Restorff (1933) showed that if all but one item of a list are similar on some dimension, memory for the dissimilar item will be enhanced. Many similar phenomena have since been reported (Wallace, 1965), and these issues have begun to attract increasing attention in neuroscience, particularly with regard to the role of the hippocampus (Nyberg, 2005). The relationship between change detection and memory formation will be a recurring theme in this thesis.

1.1.4 Forgetting in human memory

The question of why we forget (or distort) our experiences has been a central question in human memory research since its inception. This question is much too big a topic to adequately address here (see Wixted, 2004, for a review), but I will give the reader a taste of some of the key issues. The initial theoretical battle lines were drawn between decay theories (which viewed forgetting as a consequence of time-dependent memory trace degradation) and interference theories (which viewed forgetting as a consequence of competition between items). Some of the earliest evidence against decay theory was reported by Jenkins and Dallenbach (1924), who showed that subjects

who slept during a retention interval exhibited improved recall compared to subjects who stayed awake. If time was really the crucial factor (as classical decay theory asserted), then one would expect no difference between these conditions; in contrast, interference theory naturally interprets the findings of Jenkins and Dallenbach, since sleep presumably prevents the acquisition of potentially interfering new memories.

Some went farther, accusing decay theory of being scientifically vacuous. McGeoch (1932) is worth quoting in full here to appreciate the full glory of his wrath:

> In scientific descriptions of nature time itself is not employed as a causative factor nor is passive decay with time ever found. In time iron, when unused, may rust, but oxidation, not time, is responsible. In time organisms grow old, but time enters only as a logical framework in which complex biochemical processes go their ways. In time all events occur, but to use time as an explanation would be to explain in terms so perfectly general as to be meaningless. As well might one use space or number. To say that mere disuse, time unfilled for the acquisitions in question, will account for forgetting is, even were the correlation perfect, to enunciate a proposition too general to be meaningful.

So decay theory appeared to be dead, but how exactly does interference cause forgetting? Do new items overwrite the memory traces of older items, or do they coexist and compete at retrieval? Clearly, this is the same question that arises repeatedly in the experimental literature discussed above.

Barnes and Underwood (1959) tried to attack this question using a procedure that came to be known as "modified modified free recall" (MMFR), in which subjects are given time (and encouraged) to recall all items that are associated with a cue. Subjects in the Barnes and Underwood study first learned a list of $A_i - B_i$ pairs, followed by varying amounts of training with a second list of $A_i - C_i$ pairs. As the amount of $A_i - C_i$ training increased, recall of B_i decreased. Under the assumption

that the MMFR procedure eliminates retrieval competition, Barnes and Underwood concluded that the only viable explanation for their results was that the B_i memory traces were being overwritten by the C_i information.

This interpretation was called into question by later studies. For example, Brown (1976) summarized studies showing that subjects tended to remember more B_i items after a delay, a phenomenon precisely analogous to spontaneous recovery in Pavlovian conditioning. Such a phenomenon appeared more consistent with a temporary inhibition of retrieval rather than permanent loss of memory.

Today, most computational theories of human memory explain forgetting in terms of retrieval competition (see Norman et al., 2006, for a review). These models typically assume that the memory system stores an indelible copy of each experience, and memory "loss" arises in the retrieval stage, when items compete for retrieval. While the explanatory reach of these models is truly impressive, they are inadequate in at least two ways. First, from a biological perspective, it seems highly unlikely that memories are stored as separate traces. The view of the hippocampus as a distributed memory system is widely accepted in the neurobiological literature (McNaughton and Morris, 1987; Rolls, 2010). If one accepts that episodic memory relies crucially on the hippocampus, then allowing memory representations to overlap in a distributed fashion is contradictory to the separate storage assumption.

Second, there are subtle arguments against separate storage from psychology. Ratcliff et al. (1990) observed that repeating items on a list tends to aid their recognition without degrading recognition of other items (the *null list-strength effect*). Shiffrin et al. (1990) argued that this finding is consistent with a model in which repetition of items results in refinement of existing traces, rather than formation of new traces.

1.2 The Bayesian perspective

Why is it so difficult to modify memories? Why are extinction and amnesia transient? While it is difficult to give an answer to these questions that encompasses all the relevant data, I will venture a theory: The rules governing the formation and modification of memories arise naturally from rational reasoning about the statistics of the environment. In other words, there are hidden variables in our environment about which our brains are constantly learning, and the memory traces that are formed as a consequence of this learning process reflect the statistical structure of these hidden variables. While this general idea has appeared in various forms over the years (e.g., Anderson, 1990; Steyvers et al., 2006), the novelty of my approach lies in the particular *representation* of hidden variables. This representation offers a simple and clear way of thinking about when new memory traces should be formed, and when old ones should be modified.

1.2.1 Latent causes

The central representation in my theory is a hidden variable I call a *latent cause*. In Chapter 3 I provide more background on earlier work upon which my own is based; for now, I will try to give the basic gist. For concreteness, let us imagine an animal in a fear conditioning experiment. During the training phase, the animal observes a series of tone-shock pairs; during the extinction phase, the animal observes the tone by itself. As I explained above, it has traditionally been thought that the animal learns an association between tone and shock over the course of training (Pearce and Bouton, 2001), which leads to the erroneous prediction that extinction results in the unlearning of the association. The latent cause model provides a very different metaphor: observational data (tone-shock or tone-alone trials) are generated by latent causes, drawn from a distribution P(data|cause). The latent causes needn't have a

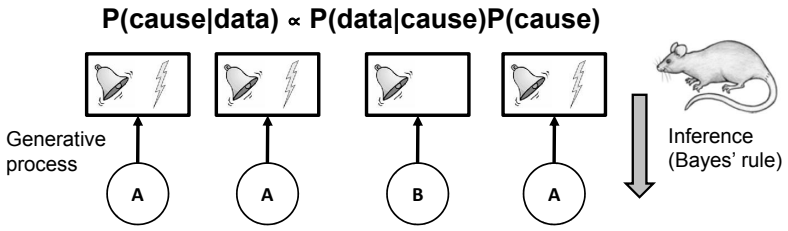

Figure 1.2: **Schematic of the latent cause theory**. Each box represents the animal's observations on a single trial. The circles represent latent causes, labeled according to their identity. The upward arrows denote probabilistic dependencies: observations are *generated* by latent causes. The animal does not get to observe the latent causes; it must infer these by inverting the generative model using Bayes' rule, as indicated by the downward arrow. As shown at the top of the schematic, Bayes' rule defines the probability of latent causes conditional on observations, which is obtained (up to a normalization constant) by multiplying the probability of observations given hypothetical causes (the likelihood) and the probability of the hypothetical latent causes (the prior).

direct physical meaning; it is better to think of them as hypothetical entities posited by the animal as a means of organizing its observational data.

Given some observational data, the animal computes the conditional distribution over possible causes given the observation—commonly known as the *posterior* distribution (shown schematically in Figure 1.2). This distribution may include previously inferred causes, as well as the hypothesis that a completely new cause generated the data. Mathematically, the posterior is given by the axiom of probability theory known as *Bayes' rule*:

$$P(\text{cause}|\text{data}) \propto P(\text{data}|\text{cause})P(\text{cause}). \tag{1.1}$$

The second term in Eq. 1.1 is known as the *prior*—it encodes the animal's "inductive bias" (Griffiths et al., 2010) about which latent causes are likely *a priori*. In later chapters, I go into much greater detail about what kinds of inductive biases the brains of humans and animals might be using.

When several observations are assigned to the same latent cause, the summary

statistics of these observations become associated with that cause. For example, when all the training trials are assigned to a single latent cause, that latent cause's distribution over observations becomes concentrated on tone-shock pairs. During extinction, this distribution is a poor predictor of tone-alone trials; because the posterior distribution must sum to 1, reducing the probability of assigning these trials to the training cause results in a corresponding increase in the probability of assigning them to a new, "extinction" cause. This is a precise formalization of the frequently proposed idea that extinction involves new learning (e.g., Bouton, 1993; Delamater, 2004). Thus, we can think of each latent cause as encoding a trace of a set of observations, and new causes are inferred when none of the previous traces are good predictors of incoming observations. The Bayesian framework provides a rational answer to the question of when a new memory should be formed. The rest of this thesis is devoted to a wide-ranging exploration of this basic idea.

1.2.2 Foundational issues

Before moving on, it is worth dwelling on the foundational logic of this framework as an empirical enterprise. Specifically, suppose we observe humans or other animals behaving in accordance with the predictions of a Bayesian model; can we conclude from this that these individuals are carrying out Bayesian computations? This is obviously an ill-posed problem, since many computations (Bayesian or non-Bayesian) could give rise to the same behavior. Even if we were to accept that the behavior arises from a "rational" analysis of the information processing problem, it is still an ill-posed problem: a basic tenet of the Bayesian philosophy is that one's posterior beliefs (rationally calculated according to the probability calculus) are always dependent on one's prior beliefs, and the prior beliefs are not dictated by any universal law. Priors are *subjective*, and hence any information processing problem in fact admits numerous solutions, all of which are rational.

This issue highlights the problematic nature of the term "rational," which is inherited from the rationalist tradition in analytical philosophy, originating with the work of Descartes, Leibniz and Spinoza. Those philosophers viewed knowledge as deriving deductively from a set of logical or mathematical principles. In the case of Bayesian inference, however, one cannot pin down a single deductively correct conclusion without first committing to a set of subjective prior beliefs. An implication of this fact is that for any given behavior we can often find a rational analysis with which it is consistent. To make this an empirical enterprise rather than an exercise in speculative philosophy, we must appeal to principles outside of the rational analysis itself.

One principle to which we can appeal might be termed "the principle of meta-rationality." Just as the brain might reason about the world using Bayesian inference, we (as scientists) can reason about the brain using the same theoretical apparatus. We have inductive biases about what kinds of models are better than others (e.g., simplicity, smoothness, Markovian structure). We can justify a preference for Bayesian explanations of behavior as one more inductive bias. Of course, this will not be satisfying to the psychologist without such a bias (Jones and Love, 2011).

We can continue to apply the principle of meta-rationality recursively, asking: where do our own inductive biases come from? For the Bayesian psychologist, one answer is evolution. Correct probabilistic reasoning, all other things being equal, is likely to increase fitness over incorrect reasoning. More importantly, correct *priors* are likely to increase fitness over incorrect priors. What this means is that if an organism's prior over some variable matches the environmental distribution of this variable, the organism is better equipped to survive and reproduce (Geisler and Diehl, 2003). Thus, if the priors used by the brain reflect environmental statistics, this favors not only Bayesian explanations in general, but *particular* Bayesian explanations.

In many situations (such as the ones analyzed in this thesis), we cannot directly

measure the relevant environmental statistics. This does not, however, invalidate the general argument: Bayesian cognitive capabilities are powerful tools for survival and reproduction, and hence ideal candidates for natural selection. Thus, from the perspective of evolutionary psychology, Bayesian models of cognition have high *a priori* probability.

Even if one rejects the foregoing arguments, there is another case to be made in favor of Bayesian models, eloquently expressed by Marr (1982):

> ... [T]rying to understand perception by studying only neurons is like trying to understand bird flight by studying only feathers: It just cannot be done. In order to understand bird flight, we have to understand aerodynamics; only then do the structure of feathers and the different shapes of birds' wings make sense (p. 27).

Marr was not talking about Bayesian models specifically, but the argument applies to all "rational" characterizations of behavior. Stringently empiricist approaches to psychology describe, but do not *explain*—they do not "make sense." Bayesian models offer a sweeping framework for sense-making. The danger is that they make sense of behavior in the same way that phlogiston made sense of combustion or ether made sense of gravity. But one reason that those concepts faded from physics is that at a certain point they no longer offered parsimonious or consistent explanations of empirical phenomena. I'm confident that the same scientific process can be used to weed out empirically recalcitrant Bayesian models.

1.3 Organization of the thesis

In Chapter 2, I provide a brief technical introduction to the probabilistic models and Bayesian methods that recur throughout the thesis.

In Chapters 3 and 4, I present two detailed implementations of the latent cause

framework and their application to phenomena in Pavlovian conditioning. In particular, I explore the circumstances in which the recovery of a conditioned response following extinction will be observed. Such recovery has long been considered evidence that extinction training does not erase the original memory, but instead creates a new memory. I formalize this idea in terms of latent causes: extinction training (under normal circumstances) leads to the inference that a new latent cause is active, thereby preventing the extinction trials from modifying the animal's beliefs about the latent cause associated with acquisition trials. Simulations demonstrate the explanatory power of the computational framework: Renewal, latent inhibition, and spontaneous recovery, as well as the dependence of recovery on numerous experimental parameters, can be accounted for by this framework. The basic lesson from these theoretical chapters is that *memories correspond to inferences about latent causes.*

My framework not only provides explanations for historical data, but has also inspired new experiments. Chapters 5-7 present new behavioral experiments in humans and rats, illustrating some of the rich inductive biases that arise from the latent cause framework (and more generally the Bayesian perspective on learning). One of these is that gradually changing the statistics between training and extinction should reduce spontaneous recovery, since the posterior probability that these phases were generated by different latent causes will be diminished. In Chapter 5, I describe two fear conditioning experiments in rats that use a "gradual extinction" paradigm (in which the CS-US contingency was gradually reduced during extinction) which essentially eliminated recovery of fear. Chapter 6 presents a conceptually similar set of findings using a visual memory task in humans. Chapter 6 also describes a variant of the latent cause model that could account for the behavioral findings.

In Chapter 7, I explore a different sort of inductive bias: the "simplicity principle" (Chater and Vitányi, 2003), which in the latent cause framework means there is a preference for explaining the observational data in terms of a parsimonious number

of latent causes.

One of the important ideas from the reconsolidation literature is that retrieving a memory renders it labile. In Chapter 8 I present an fMRI study in humans that investigates this idea in a human list-learning paradigm developed by Hupbach and colleagues (Hupbach et al., 2007, 2008, 2009, 2011; Jones et al., 2012) which uses reminders to induce source memory misattributions. In Sederberg et al. (2011), my colleagues and I proposed that a key determinant of memory misattributions is the reinstatement of "mental context" hypothesized by a number of human memory models, in particular the Temporal Context Model (TCM; Howard and Kahana, 2002; Sederberg et al., 2008; Gershman et al., 2012). I provide direct neural evidence of the hypothesized neural context reinstatement, and show that it is predictive of which items will be subsequently misattributed, as well as the parametric degree of confidence in the misattribution. I discuss ways in which the mechanistic ideas underlying TCM (and supported by my brain imaging data) may tie into the computational-level ideas embodied by the latent cause theory.

Finally, in Chapter 9, I summarize the main lessons learned from this work, and provide an outlook for the future.

Chapter 2

Probabilistic models and Bayesian methods

This chapter provides a brief introduction to the probabilistic models and Bayesian methods used throughout the thesis. First, I introduce mixture models and their Bayesian nonparametric ("infinite") extensions. These are the basic building blocks of the models presented in later chapters. I then describe Monte Carlo methods for approximating Bayesian inference in these models, focusing on the sequential Monte Carlo algorithm known as particle filtering. Some of the material in this chapter was adapted from Gershman and Blei (2012).

2.1 Mixture models and clustering

In a mixture model, each observed data point is assumed to belong to a cluster. In posterior inference, we infer a grouping or clustering of the data under these assumptions—this amounts to inferring both the identities of the clusters and the assignments of the data to them. Mixture models are used for understanding the group structure of a data set and for flexibly estimating the distribution of a population.

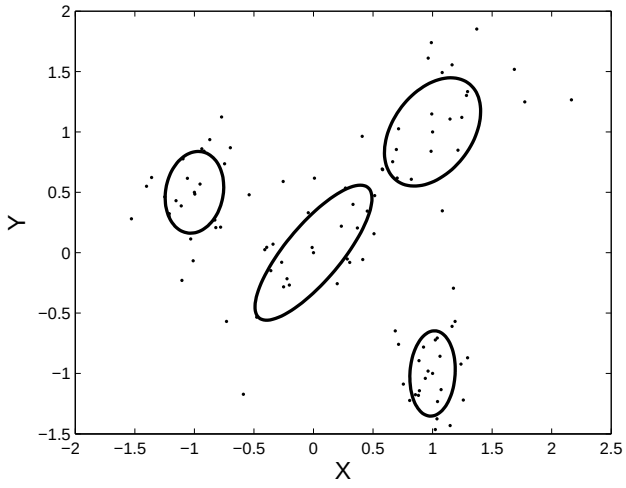

Figure 2.1: **Draws from a Gaussian mixture model**. Ellipses show the standard deviation contour for each mixture component. Reprinted from Gershman and Blei (2012).

2.1.1 Finite mixture modeling

A finite mixture model assumes that there are K clusters, each associated with a parameter θ_k. Each observation y_n is assumed to be generated by first choosing a cluster c_n according to $P(c_n)$ and then generating the observation from its corresponding observation distribution parameterized by θ_{c_n}. Each cluster specifies a hypothetical distribution $F(y_n|\theta_{c_n})$ over the observed data.

Finite mixtures can accommodate many kinds of data by changing the data generating distribution. For example, in a Gaussian mixture model the data—conditioned on knowing their cluster assignments—are assumed to be drawn from a Gaussian distribution. The cluster parameters θ_k are the means of the components (assuming known variances). Figure 2.1 illustrates data drawn from a Gaussian mixture with four clusters.

Bayesian mixture models further contain a prior over the mixing distribution $P(c)$, and a prior over the cluster parameters: $\theta \sim G_0$. (We denote the prior over cluster parameters G_0 to later make a connection to Bayesian nonparametric mixture mod-

els.) In a Gaussian mixture, for example, it is computationally convenient to choose the cluster parameter prior to be Gaussian. A convenient choice for the distribution on the mixing distribution is a Dirichlet. We will build on fully Bayesian mixture modeling when we discuss Bayesian nonparametric mixture models.

This generative process defines a joint distribution over the observations, cluster assignments, and cluster parameters,

$$P(\mathbf{y}, \mathbf{c}, \theta) = \prod_{k=1}^{K} G_0(\theta_k) \prod_{n=1}^{N} F(y_n|\theta_{c_n}) P(c_n), \tag{2.1}$$

where the observations are $\mathbf{y} = \{y_1, \ldots, y_N\}$, the cluster assignments are $\mathbf{c} = \{c_1, \ldots, c_N\}$, and the cluster parameters are $\theta = \{\theta_1, \ldots, \theta_K\}$. The product over n follows from assuming that each observation is conditionally independent given its latent cluster assignment and the cluster parameters.

Given a data set, we are usually interested in the cluster assignments, i.e., a grouping of the data.[1] We can use Bayes' rule to calculate the posterior probability of assignments given the data:

$$P(\mathbf{c}|\mathbf{y}) = \frac{P(\mathbf{y}|\mathbf{c})P(\mathbf{c})}{\sum_{\mathbf{c}} P(\mathbf{y}|\mathbf{c})P(\mathbf{c})}, \tag{2.2}$$

where the likelihood is obtained by marginalizing over settings of θ:

$$P(\mathbf{y}|\mathbf{c}) = \int_{\theta} \left[\prod_{n=1}^{N} F(\mathbf{y}|\theta_{c_n}) \prod_{k=1}^{K} G_0(\theta_k) \right] d\theta. \tag{2.3}$$

A G_0 that is conjugate to F allows this integral to be calculated analytically. For example, the Gaussian is the conjugate prior to a Gaussian with fixed variance, and this is why it is computationally convenient to select G_0 to be Gaussian in a mixture

[1] Under the Dirichlet prior, the assignment vector $\mathbf{c} = [1, 2, 2]$ has the same probability as $\mathbf{c} = [2, 1, 1]$. That is, these vectors are equivalent up to a "label switch." Generally we do not care about what particular labels are associated with each class; rather, we care about *partitions*—equivalence classes of assignment vectors that preserve the same groupings but ignore labels.

of Gaussians model.

The posterior over assignments is intractable because computing the denominator (marginal likelihood) requires summing over every possible partition of the data into K groups. (This problem becomes more salient in the next section, where we consider the limiting case $K \to \infty$.) We can use Monte Carlo methods to approximate this computation. These methods are discussed further below.

2.1.2 The Chinese restaurant process

When we analyze data with the finite mixture of Equation 2.1, we must specify the number of latent clusters (e.g., hypothetical cognitive processes) in advance. In many data analysis settings, however, we do not know this number and would like to learn it from the data. Bayesian nonparametric clustering addresses this problem by assuming that there is an infinite number of latent clusters, but that a finite number of them is used to generate the observed data. Under these assumptions, the posterior provides a distribution over the number of clusters, the assignment of data to clusters, and the parameters associated with each cluster. Furthermore, the predictive distribution, i.e., the distribution of the next data point, allows for new data to be assigned to a previously unseen cluster.

The Bayesian nonparametric approach finesses the problem of choosing the number of clusters by assuming that it is infinite, while specifying the prior over infinite groupings $P(\mathbf{c})$ in such a way that it favors assigning data to a small number of groups. The prior over groupings is called the *Chinese restaurant process* (CRP; Aldous, 1985; Pitman, 2002), a distribution over infinite partitions of the integers; this distribution was independently discovered by Anderson (1991) in the context of his rational model of categorization. Variants of this prior have been widely used in cognitive science to model probabilistic reasoning about combinatorial objects of unbounded cardinality (Anderson, 1991; Collins and Koechlin, 2012; Gershman et al.,

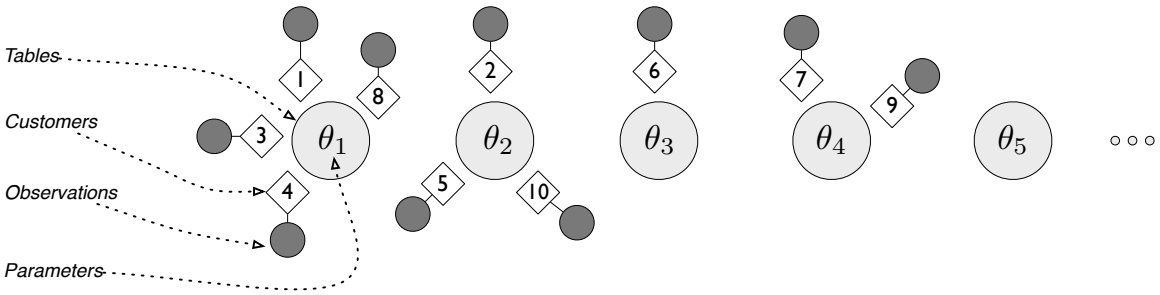

Figure 2.2: **The Chinese restaurant process**. The generative process of the CRP, where numbered diamonds represent customers, attached to their corresponding observations (shaded circles). The large circles represent tables (clusters) in the CRP and their associated parameters (θ). Note that technically the parameter values $\{\theta\}$ are not part of the CRP *per se*, but rather belong to the full mixture model. Reprinted from Gershman and Blei (2012).

2010; Goldwater et al., 2009; Sanborn et al., 2010).

The CRP derives its name from the following metaphor. Imagine a restaurant with an infinite number of tables,[2] and imagine a sequence of customers entering the restaurant and sitting down. The first customer enters and sits at the first table. The second customer enters and sits at the first table with probability $\frac{1}{1+\alpha}$, and the second table with probability $\frac{\alpha}{1+\alpha}$, where α is a positive real. When the nth customer enters the restaurant, she sits at each of the occupied tables with probability proportional to the number of previous customers sitting there, and at the next unoccupied table with probability proportional to α. At any point in this process, the assignment of customers to tables defines a random partition. A schematic of this process is shown in Figure 2.2.

More formally, let c_n be the table assignment of the nth customer. A draw from this distribution can be generated by sequentially assigning observations to classes

[2]The Chinese restaurant metaphor is due to Pitman and Dubins, who were inspired by the seemingly infinite seating capacity of Chinese restaurants in San Francisco.

with probability

$$P(c_n = k|\mathbf{c}_{1:n-1}) \propto \begin{cases} \frac{m_k}{n-1+\alpha} & \text{if } k \leq K_+ \text{ (i.e., } k \text{ is a previously occupied table)} \\ \frac{\alpha}{n-1+\alpha} & \text{otherwise (i.e., } k \text{ is the next unoccupied table)} \end{cases} \tag{2.4}$$

where m_k is the number of customers sitting at table k, and K_+ is the number of tables for which $m_k > 0$. The parameter α is called the *concentration parameter*. Intuitively, a larger value of α will produce more occupied tables (and fewer customers per table).

2.1.3 Chinese restaurant process mixture models

The Bayesian nonparametric clustering model uses the CRP in an infinite-capacity mixture model (Antoniak, 1974; Anderson, 1991; Rasmussen, 2000). Each table k is associated with a cluster and with a cluster parameter θ_k, drawn from a prior G_0. We emphasize that there are an infinite number of clusters, though a finite data set only exhibits a finite number of active clusters. Each data point is a "customer," who sits at a table c_n and then draws its observed value from the distribution $F(y_n|\theta_{c_n})$. The concentration parameter α controls the prior expected number of clusters (i.e., occupied tables) K_+. In particular, this number grows logarithmically with the number of customers N: $\mathbb{E}[K_+] = \alpha \log N$ (for $\alpha < N/\log N$).

By examining the posterior over partitions, we can infer both the assignment of observations to clusters and the number of clusters. In addition, the (approximate) posterior provides a measure of confidence in any particular clustering, without committing to a single cluster assignment. Notice that the number of clusters can grow as more data are observed. This is both a natural regime for real-world data, and it makes the CRP mixture robust to new data that is far away from the original observations.

When we analyze data with a CRP, we form an approximation of the joint poste-

rior over all latent variables and parameters. In practice, there are two uses for this posterior. One is to examine the likely partitioning of the data. This gives us a sense of how are data are grouped, and how many groups the CRP model chose to use. The second use is to form predictions with the posterior predictive distribution. With a CRP mixture, the posterior predictive distribution is

$$P(y_{n+1}|\mathbf{y}_{1:n}) = \sum_{\mathbf{c}_{1:n+1}} \int_{\theta} P(y_{n+1}|c_{n+1},\theta)P(c_{n+1}|\mathbf{c}_{1:n})P(\mathbf{c}_{1:n},\theta|\mathbf{y}_{1:n})d\theta. \tag{2.5}$$

Since the CRP prior, $P(c_{n+1}|\mathbf{c}_{1:n})$, appears in the predictive distribution, the CRP mixture allows new data to possibly exhibit a previously unseen cluster.

2.2 Monte Carlo methods and particle filtering

Recall that for the mixture models described above (and for many other models of interest), exactly computing Bayes' rule is intractable, a consequence of the denominator (the marginal likelihood) being a sum of exponentially many terms. A very general and flexible approach to approximating Bayes' rule is using Monte Carlo methods. The essential idea is to approximate the true posterior with a set of M samples, which we denote by $\mathbf{c}^{(1:M)}$:

$$P(\mathbf{c}|\mathbf{y}) \approx \frac{1}{M}\sum_{m=1}^{M} \delta\left[\mathbf{c}, \mathbf{c}^{(m)}\right], \tag{2.6}$$

where $\delta[\cdot,\cdot] = 1$ if its arguments are equal, and 0 otherwise. Samples can also be used to approximate expectations. Letting $f(\mathbf{c})$ denote a function of the latent variables,

$$\mathbb{E}_{P(\mathbf{c}|\mathbf{y})}[f(\mathbf{c})] \approx \frac{1}{M}\sum_{m=1}^{M} f(\mathbf{c}^{(m)}). \tag{2.7}$$

As long as $\mathbf{c}^{(m)} \sim P(\mathbf{c}|\mathbf{y})$, the Monte Carlo expectation will converge to the true expectation as $M \to \infty$ by the strong law of large numbers.

There are a number of computational challenges in applying Monte Carlo methods. First, we may not be able to directly draw samples from $P(\mathbf{c}|\mathbf{y})$. One approach is to draw samples from a Markov chain whose stationary distribution is the posterior; this is the basis of *Markov chain Monte Carlo* algorithms (Robert and Casella, 2004). Another approach is to draw samples from a *proposal distribution* $Q(\mathbf{c})$ and then calculate a weighted Monte Carlo approximation:

$$P(\mathbf{c}|\mathbf{y}) \approx \sum_{m=1}^{M} w_m \delta\left[\mathbf{c}, \mathbf{c}^{(m)}\right], \tag{2.8}$$

where $w_m \propto P(\mathbf{c}, \mathbf{y})/Q(\mathbf{c})$. This is known as *importance sampling*. Again, the strong law of large numbers guarantees that expectations of functions under this approximation will approach their true value as $M \to \infty$.

A second challenge is that for large data sets it may be inefficient to sample from the entire set of latent variables; instead, we may wish to approximate the *filtering* distribution,

$$P(c_n|\mathbf{y}_{1:n}) = \sum_{\mathbf{c}_{1:n-1}} P(c_n|\mathbf{y}_{1:n}, \mathbf{c}_{1:n-1}) P(\mathbf{c}_{1:n-1}|\mathbf{y}_{1:n-1}). \tag{2.9}$$

The filtering distribution represents the posterior belief about the latent variable at time n after marginalizing the latent variables at the previous time steps.

Particle filtering (Doucet et al., 2001) is a variant of importance sampling that approximates the filtering distribution. The basic scheme at time n is as follows:

1. Draw M samples from a proposal distribution $Q(c_n|\mathbf{c}_{1:n-1}^{(1:M)})$.
2. Calculate the importance weights: $w_m \propto w_{m-1} P(c_n|y_n)/Q(c_n|\mathbf{c}_{1:n-1}^{(1:M)})$.

This is simply a recursive application of importance sampling that approximates the

marginalization in Eq. 2.9 with a set of samples (aka particles). In some versions of the algorithm, the particles are resampled (with replacement) after each iteration according to the probability distribution defined by $\mathbf{w}_{1:M}$. The optimal proposal distribution for particle filtering is $P(c_n|c_{n-1}, y_n)$; for some models (e.g., in Chapter 3), the optimal proposal distribution can be computed analytically.

Chapter 3

Context, learning and extinction: a latent cause theory

An enduring problem in the study of Pavlovian conditioning is how animals learn about the causal structure of their environment (Blaisdell et al., 2006). Most theories frame conditioning as the learning of associations between stimuli and reinforcement (Rescorla and Wagner, 1972; Pearce and Bouton, 2001). Under a statistical interpretation, these associations are parameters of a generative model in which stimuli cause reinforcement (Kakade and Dayan, 2002). However, evidence suggests that animals may employ more flexible models, learning, for example, that some stimuli are causally unrelated to reinforcement (Dayan and Long, 1998; Dayan et al., 2000). A more radical departure are *latent cause* models (Courville et al., 2002, 2004; Courville, 2006), in which both stimuli and reinforcement are attributed to causes that are hidden from observation. One motivation for such models is the finding that learned relationships between cues and reinforcement are not necessarily erased following extinction: returning the animal to the original training context after extinction in a different context can lead to renewal of the conditioned response (Bouton and Bolles,

[0]The work described in this chapter has been published as S.J. Gershman, D.M. Blei, & Y. Niv (2010). Context, learning, and extinction. *Psychological Review, 117,* 197–209.

1979a; Bouton, 2004). These and related data can be characterized by a latent cause model in which different latent causes are associated with the training and extinction contexts.

One problem with latent cause models is that the number of different latent causes is in general unknown. The challenge, then, is to formulate a learning algorithm that can infer new causes as it gathers observations, as well as learn the statistical relationships between causes and observations. Recently, Redish et al. (2007) have formulated such a theory of extinction learning, combining the well-studied framework of temporal difference reinforcement learning (Sutton and Barto, 1998; Schultz et al., 1997) with a state classification mechanism that allows the state space to expand adaptively. In their model, states can be loosely interpreted as latent causes, serving to explain both stimuli and reinforcement in terms of an underlying discrete variable.

In this chapter, I suggest a new model of latent cause inference in animal learning that is based in a normative statistical framework. With this model, I address certain limitations of the theory of Redish et al. (2007), while still capturing its essential insights. The model agrees with their assertion that the computational problem the animal must solve is one of structure learning. I posit that the computational principles at work in structure learning are based on a generative model of the environment that specifies the animal's *a priori* beliefs about how its observations were generated by latent causes. Given a set of observations, the problem facing the animal is to combine its prior beliefs with the evidence provided by the observations to infer which causes were in action. At the algorithmic level, I identify several features of Redish et al.'s model that make it difficult to account for relevant data and show how these are obviated in my model. Finally, drawing on a suggestion by Redish et al., I make explicit the computational role played by the hippocampus in my model and use this to explain developmental changes in learning.

3.1 Redish et al.'s model

The data motivating the model of Redish et al. (2007) come from a conditioning procedure studied by Bouton and Bolles (1979a): In the acquisition phase, the animal is placed in context A and exposed in each trial to both a stimulus cue and a reinforcer; eventually the cue comes to evoke a conditioned response. In the extinction phase, the animal is then placed in a new context (B) and exposed in each trial to the cue in the absence of reinforcement, until the cue ceases to evoke the conditioned response. It would appear, at first glance, that the animal has "unlearned" its response to the cue. However, if the animal is returned to the original context (A) in a test phase and presented with the cue, the response is restored, strongly suggesting otherwise. Rather, it seems that the animal has learned a new relationship between the cue and the reinforcer during extinction that was somewhat limited to the context B.

This phenomenon, known as "ABA renewal," is explained by Redish et al.'s model in terms of two simultaneous processes: a value learning process and a state classification process. The first updates values associated with states (and potentially actions), using a form of the temporal difference learning algorithm (Sutton and Barto, 1998), which is closely related to the Rescorla-Wagner update rule (Rescorla and Wagner, 1972). A state's value represents a prediction about future reinforcement in that state. The temporal difference learning rule incrementally updates values in proportion to the discrepancy between predicted and received reinforcement (the "prediction error"). In the Pavlovian version of the renewal paradigm described above, the animal's conditioned response is presumed to be proportional to the current state's value (Dayan et al., 2006). In the operant version modeled by Redish et al. (2007), the probability of the animal taking a particular action (e.g., lever press) is proportional to the state-action value.

The innovation of Redish et al. is to introduce a state classification process that determines what state the animal is currently in and creates new states when the ob-

servation statistics change. The observations, in this case, are defined to be tuples consisting of {context, stimulus, immediate reinforcement, time since last reinforcement}. The actual mechanics of the state classification are quite sophisticated, and I refer the reader to the original paper. In brief, a competitive learning model using radial basis functions and classifier expansion (Hertz et al., 1991; Grossberg, 1976) partitions the input space into multivariate Gaussian state prototypes; temporal difference learning then operates on these states. For present purposes, the important aspect of this process is that each state is associated with an observation prototype, and a new observation is classified as a particular state on the basis of its match to the state's prototype. When the observation fails to match any prototype (as determined by a threshold), a new state/prototype is inferred. A local estimate of the average negative prediction error modulates this process: when prediction errors are tonically negative, a new state is more likely to be inferred.

According to this model, acquisition in the ABA renewal paradigm proceeds according to the value learning process, with all training observations being assigned to the same state (since their statistics are homogeneous). During extinction, the absence of the predicted reinforcement results in tonic negative prediction errors. Combined with a context change, this results in the state classification process creating a new state. Thus one state is associated with reinforcement, and another state is associated with no reinforcement. When the animal is returned to the training context, it identifies its observations as belonging to the state associated with reinforcement (on the basis of the contextual cue), and therefore produces the conditioned response.

One implication of this model is that new states are unlikely to be inferred when prediction errors are tonically *positive*. Evidence in contradiction of this hypothesis comes from the context-dependence of latent inhibition (Hall and Honey, 1989). The latent inhibition procedure is, in some sense, a concomitant manipulation to

extinction: an animal is first exposed to a stimulus in the absence of reinforcement, and later conditioned with pairing of the stimulus and reinforcement. In this case, animals are slower to acquire a conditioned response, as compared to animals that have not been pre-exposed. However, if the pre-exposure and conditioning phases are conducted in different contexts, the latent inhibition effect is diminished. Here the conditioning phase is accompanied by positive prediction errors (as the reinforcement is unexpected following the pre-exposure) which, according to Redish et al.'s model, should not result in the inference of a new state. Hence, their model mispredicts that a shift in context will not affect latent inhibition.[1] This problem was also noted by Redish et al. (2007).

Another problem is that because the values associated with new states are initialized to 0, Redish et al.'s model does not predict ABC and AAB renewal (Bouton and Bolles, 1979a; Bouton and King, 1983), in which the test trials occur in a completely new context. In both these cases, conditioned responding returns during the test phase. This can be fairly easily accommodated by initializing the values of new states to some prior belief about state values, as I will discuss below in connection with my model.

Apart from these problems, the idea of state classification on the basis of observation statistics is a fundamental contribution. In formulating a quantitative theory of how animals solve this problem, my goal is to understand the statistical principles underlying this insight. To this end, I propose a new model that is conceptually aligned with that of Redish et al., but more directly descended from the latent cause theory of Courville (2006).

The rest of this chapter is organized as follows. I first describe an infinite-capacity mixture model and a particle filter algorithm for performing inference in this model. I

[1]Redish et al.'s model will demonstrate latent inhibition if the modulation of state classification by tonic prediction error is weak. In this case, state classification is driven primarily by the match between the current observation and the prototypes. However, this scenario is at odds with the central role played by tonic prediction error in Redish et al.'s model.

then present the results of simulations of latent inhibition and renewal paradigms, including developmental and hippocampal manipulations. In the discussion, I compare my model to that of Redish et al. (2007), as well as several other models. Finally, I discuss some limitations of my model and propose directions for its future development.

3.2 A new model: statistical inference in an infinite-capacity mixture model

The central claim of my account is that, in order to adaptively predict events in their environment, animals attempt to partition observations into separate groups on the basis of their properties. This task is known as "clustering" in computer science and statistics and hence I will call these groups "clusters." I assume that the animal's goal is to assign observations to clusters such that the clusters correspond to different latent causes. Renewal can then be understood as the result of this clustering process. Specifically, I suggest that the animal learns to partition its observations on the basis of their features into two distinct clusters, corresponding to the acquisition and extinction phases (which are implicitly the "causes" of the animal's observations).

My basic approach is to first formulate a set of assumptions that are imputed to the animal, and then to describe how, based on these assumptions, the animal can make rational inferences about latent causes given a set of observations. The set of assumptions constitutes the *generative model*, which represents the animal's prior beliefs about the statistical structure and probabilistic dependencies between variables (both hidden and observed) in the environment. The generative model expresses the state of the animal's beliefs *prior* to making any observations. Given a set of observations, we expect the animal's beliefs (or inference) about the actual causes of these observations to change. In particular, this new state of knowledge

is expressed by a posterior distribution over unobserved (hidden) variables given the observed variables. I shall refer to this as the animal's *inference model*.[2] In the context of the classical conditioning data that I model, the inference model represents the animal's beliefs about the latent causes of its observations.

3.2.1 Generative model

I assume that the animal's observation on trial t takes the form of a discrete-valued multi-dimensional vector $\mathbf{f}_t$, with the following dimensions: reinforcement ($f_{t,1}$), cue ($f_{t,2}$) and context ($f_{t,3}$). The reinforcement dimension represents a binary unconditional stimulus delivered to the animal, e.g. $f_{t,1} \in \{\texttt{reinforcement}, \texttt{no reinforcement}\}$. The cue dimension represents a typical Pavlovian cue (or its absence). e.g., $f_{t,2} \in \{\texttt{tone}, \texttt{no tone}\}$.[3] The context dimension is an abstraction of the context manipulations typical in renewal paradigms (e.g., box color, shape, odor, etc.), which I simplify into discrete values: $f_{t,3} \in \{\texttt{context A}, \texttt{context B}, \texttt{context C}\}$.

The generative model I impute to the animal is one in which, on each trial, a single latent cause is responsible for generating all the observation features (reinforcement, cue, context). In such a *mixture model*, each trial is assumed to be generated by first sampling a cause c_t (from a known set of causes) according to a mixing distribution $P(c)$, and then sampling observation features conditioned on the cause from an observation distribution $P(\mathbf{f}|c_t)$. This type of generative model is a reasonable prior belief for many environments. In fact, it correctly expresses, to a first approximation, the process by which many conditioning procedures are generated: first a phase (e.g., conditioning, extinction, test) is selected, and then a set of stimuli are selected conditioned on the phase. If the animal assumes that each observation is probabilistically generated by a single latent cause, then clustering is the process of recovering these

[2] This is also sometimes referred to as the *recognition model* (Dayan and Abbott, 2001).

[3] The choice of discrete-valued observations is not crucial to my formalism; I have used real-valued features and obtained essentially the same results.

causes on the basis of its observations.[4]

The mixture model described so far implicitly assumes that the animal knows how many possible causes there are in the environment. This seems an unreasonable assumption about the structure of the animal's environment, as well as the animal's *a priori* knowledge about its observations. Furthermore, as I will discuss later, there is evidence that animals can flexibly infer the existence of new causes as more observations are made. I thus use a generative model that allows for an unbounded (expanding) number of latent causes (an *infinite capacity* mixture model, as described below). In this model, the animal prefers a small number of causes but can, at any time, infer the occurrence of a new latent cause when the data support its existence and thus decide to assign its current observation to a completely new cluster.

Formally, let us denote a partition of observations (trials) $1, \ldots, t$ by the vector $\mathbf{c}_{1:t}$. A partition specifies which observations were generated by which causes, such that $c_t = k$ indicates that the observation t was assigned to cluster k. In my model, the animal's prior over partitions is the CRP introduced in Chapter 2, which generates cause k with probability

$$P(c_{t+1} = k|\mathbf{c}_{1:t}) = \begin{cases} \frac{N_k}{t+\alpha} & \text{if } k \leq K_t \text{ (i.e., } k \text{ is an old cause)} \\ \frac{\alpha}{t+\alpha} & \text{if } k = K_t + 1 \text{ (i.e., } k \text{ is a new cause),} \end{cases} \tag{3.1}$$

where N_k is the number of observations already generated by cause k (by default it is assumed that $c_1 = 1$) and K_t is the number of unique causes generated for observations 1 to t. The number of causes generating observations $1, \ldots, t$ is now a random variable, and can be any number from 1 to t. The concentration parameter α specifies the animal's prior belief about the number of causes in the environment.

[4]I will use the term *cause* in connection with the generative model, and the term *cluster* in connection with the inference procedure. The clusters inferred by the animal may not be identical to the true causes of its observations.

When $\alpha = 0$, the animal assumes that all observations are generated by a single cause; when α approaches ∞, the animal assumes that each observation is generated by a unique cause. In general, for $\alpha < \infty$, the animal assumes that observations will tend to be generated by a small number of causes.

The animal further assumes that once a cause has been sampled for a trial, an observation is sampled from an observation distribution conditional on the cause. Each cause is associated with a multinomial observation distribution over features, parameterized by ϕ, where $\phi_{i,j,k}$ is the probability of observing value j (e.g., `reinforcement`) for feature i given latent cause k. A common assumption in mixture models (which I adopt here) is that, in the generative model, each feature is conditionally independent of all the other features given a latent cause and the multinomial parameters. For instance, a latent cause that can be labelled as 'training trial' might generate a cue with probability $\phi_{2,tone,k=`training'} = 1$ and, independently, generate reinforcement with some probability $\phi_{1,reinforcement,k=`training'}$ (possibly less than 1), while a latent cause labeled as 'extinction trial' might generate a cue with probability 1 and reinforcement with probability 0. The conditional independence assumption expresses the idea that, given the identity of the latent cause, cues and reinforcement are separately generated, each according to its associated probability $\phi_{i,j,k}$.

I shall assume that the multinomial parameters themselves are drawn from a Dirichlet distribution (the conjugate prior for the multinomial distribution). This prior expresses the animal's predictions about the experiment before it has made any observations. Given that the animal is unlikely to have strong *a priori* predictions about the experiment before it has begun, I chose the parameters of the Dirichlet distribution so that all possible multinomial parameters have equal probability under the prior. Note that each cause is endowed with its own multinomial distribution over features; this allows different causes to be associated with different observation statistics. Every time a new cause is created by Equation 3.1, the parameters of its

corresponding multinomial distribution are drawn from the Dirichlet prior.

It may at first appear odd that causes in Equation 3.1 are generated purely on the basis of how many times a particular cause was generated in the past, and that features are generated independently from one another given a cause and multinomial parameters. Intuitively, one would expect that similar observations would be generated by the same cause. Indeed, this intuition is faithfully embodied in the model. An important point to keep in mind is that in the generative model, observations will be similar *because* they were generated by the same cause. Similarly, features will exhibit correlations because they are coupled by a common cause (e.g., the latent cause associated with the training phase of a conditioning experiment will tend to generate both the cue and the reinforcement). When faced with uncertainty about the latent causes of its observations, both of these properties will influence the animal's beliefs in the inference model (described in the next section). First, the animal will use the similarity between trials to infer what latent cause they came from. As a result, the belief about the causes of one trial will no longer be independent of the other trials. Second, the animal's beliefs about the future value of one feature (e.g., reinforcement) will depend in the inference model on its knowledge about other features (e.g., context and cue). In other words, observation features will be conditionally dependent when the latent cause is unknown (that is, in all realistic scenarios).

3.2.2 Inference

There are two components to the inference problem facing the animal: identifying the latent causes of its observations, and predicting reinforcement given a partial observation (context and cues). Because in my model prediction depends on inferences about latent causes, I address each of these in turn.

Denote the observations in trials $1, \ldots, t$ by the matrix $\mathbf{F}_{1:t}$. Given a set of observations up to trial t, what are the animal's beliefs about the latent causes of these

observations? According to Bayesian statistical inference, these beliefs are represented by the posterior distribution over partitions given the observations:

$$P(\mathbf{c}_{1:t}|\mathbf{F}_{1:t}) = \frac{P(\mathbf{F}_{1:t}|\mathbf{c}_{1:t})P(\mathbf{c}_{1:t})}{P(\mathbf{F}_{1:t})}. \tag{3.2}$$

Exact computation of the posterior in this model is computationally demanding. Moreover, for such a model to be plausibly realized by animals, learning and inference must be incremental and online. One approximate inference algorithm which is both tractable and incremental is the *particle filter* (Fearnhead, 2004) described in Chapter 2. This algorithm approximates the posterior distribution over partitions with a set of weighted samples, and has been used successfully to model a number of learning phenomena (Sanborn et al., 2006; Daw and Courville, 2008; Brown and Steyvers, 2009). The essential idea in particle filtering is to create a set of m hypothetical particles, each of which is a specific partition of all the trials into causes, and then weight these particles by how likely they are to have generated the particular set of observations that has been seen. The weights will depend on factors such as whether similar observations are clustered together in a particular particle and the number of latent causes in the partition. They will also depend on multiplicative interactions between features, such that a particle will receive larger weight to the extent that it predicts consistent *configurations* of feature values. A detailed description of the particle filter algorithm can be found in the Appendix.

I assume that the animal's general goal in a classical conditioning experiment is to predict the probability of reinforcement, when observing a "test" observation vector that lacks the first feature (i.e., where it is not yet specified whether reinforcement will or will not occur). This prediction can rely on the presence or absence of the other features (context and cue) as well as all of the animal's previous experience. In my model, this prediction is accomplished by augmenting each particle with a cluster

assignment of the test observation and then averaging the probability of reinforcement over all the particles, weighted by the posterior probability of the test cluster assignment (see Appendix for the corresponding equations). I assume that the animal's conditioned (Pavlovian) response is proportional to the predicted probability of reinforcement (Dayan et al., 2006) and so report the reinforcement prediction in the results.

3.3 Results

Except where otherwise mentioned, for the simulations reported here I used uniform Dirichlet priors over all features and $\alpha = 0.1$ as the concentration parameter.[5] All the simulations used 3000 particles.[6] For each phase (pre-exposure, conditioning, extinction), trials were identical replicas of each other (i.e., there was no noise injected into the observations). Although the output of the particle filter is stochastic (due to the sample-generating process), it returns effectively the same results on multiple runs by averaging over a large number of particles.

3.3.1 Renewal

Figure 3.1a shows experimental data from a renewal paradigm (Bouton and Bolles, 1979a) in which rats were given training in context A, extinction in context B, and then tested in either the training context (A), extinction context (B) or a novel context (C). The conditioned response measured at test was in this case conditioned suppression (but similar results have been obtained with many other preparations; see Bouton, 2004). Conditioned responding was renewed both in the training context

[5]Although α can be learned straightforwardly with the particle filter, my simulations suggest that this added flexibility does not change the results substantially, so I have fixed it to a constant value.

[6]I used a large number of particles to accurately approximate the posterior. For this reason, other inference algorithms, such as Gibbs sampling, will produce effectively the same predictions.

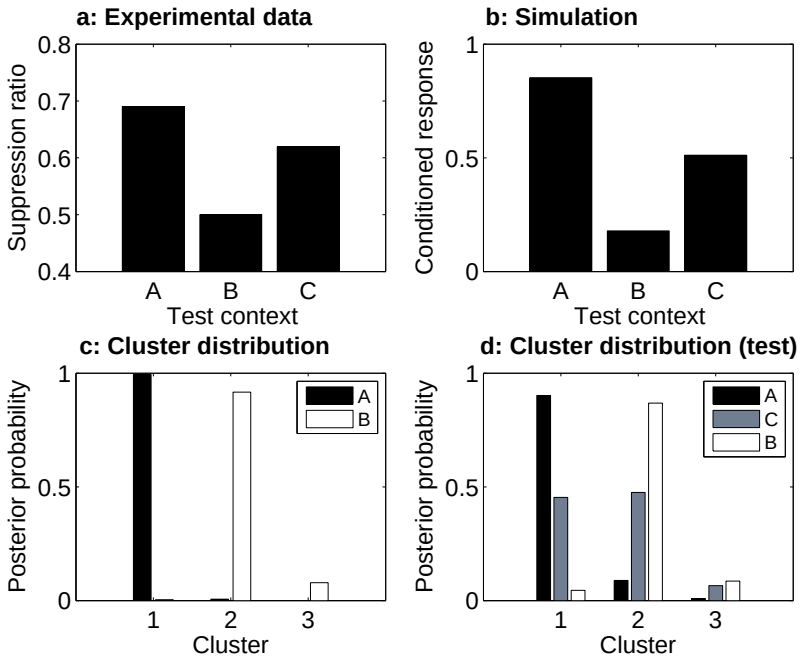

Figure 3.1: **Renewal**. Experimental (a) and simulated (b) conditioned responding to a stimulus during a test phase after conditioning and extinction. In all plots, experimentally-observed conditioned responses are plotted using their original measurement units. (a) Both returning the subject to the conditioning context and placing it in a novel context result in renewal of conditioned responding. Data replotted from (Bouton and Bolles, 1979a). (b) Simulated conditioned responding during test in the conditioning context A, extinction context B and a novel context C. (c) Posterior distribution of cluster assignments after conditioning in context A and extinction in context B. Conditioning and extinction trials tended to be assigned to different clusters, as evidenced by different modes of the posterior in the two phases. (d) Posterior distribution of cluster assignments on the first test trial in contexts A, B and C.

(ABA renewal) and in the novel context (ABC renewal), but not in the extinction context.

Figure 3.1c-d shows the results of simulating my model with conditioning in context A ($\mathbf{f} = [\texttt{reinforcement}, \texttt{tone}, \texttt{A}]$ for 20 trials), extinction in context B ($\mathbf{f} = [\texttt{no reinforcement}, \texttt{tone}, \texttt{B}]$ for 50 trials) and testing in either A ($\mathbf{f} = [?, \texttt{tone}, \texttt{B}]$), B ($\mathbf{f} = [?, \texttt{tone}, \texttt{B}]$) or C ($\mathbf{f} = [?, \texttt{tone}, \texttt{C}]$), demonstrating that my model replicates the ABA and ABC renewal effects. Similar in spirit to Redish et al.'s model, my model predicts ABA renewal as a consequence of the animal's inference that different latent

causes are active during conditioning and extinction. When the animal is returned to the conditioning context in the test phase, it infers (due to the presence of contextual cues) that the first latent cause is once again active. Because trials with the same latent cause have similar properties, the animal predicts that reinforcement is likely to occur on the test trials, and therefore emits the conditioned behavior. Thus, the importance of context in my theory derives from its usefulness in disambiguating the latent causes of observations (see also Bouton, 1993).

ABA renewal is observed in my model to the extent that a test trial matches (in its observation features) trials from the conditioning phase more than trials from the extinction phase. ABC renewal may be observed in at least three different scenarios. If C is substantially different from A or B, such that a new cluster is created, ABC renewal will be observed to the extent that the prior expectation of reinforcement in a new cluster is greater than zero. If C is not different enough to warrant a new cluster, ABC renewal may still be observed if C is equally similar to A and B, so that it gets assigned in equal proportions to their associated clusters. Yet another possibility is that when there are many more A trials than B trials, the C observation will be assigned to the cluster associated with A due to Equation 3.1 (which in the inference model will tend to assign observations to more popular clusters). Although the simulations presented here manifest the second scenario (in which trials in C are associated equally with the training cluster and the extinction cluster) I note that different parameterizations (particularly the value of α) or feature representations may result in the first scenario, in which a new cluster is inferred, which would also lead to renewal.

When the animal is tested in the same context as the extinction phase, no renewal is observed (Figure 3.1). Similarly, no renewal is observed when all three phases take place in the same context (results not shown). These results follow from the model's prediction that the same latent cause is active during extinction and test, and hence

predicts the absence of reward.

Further insight into the mechanisms underlying renewal in my model can be gained by examining the posterior distribution of clusters, shown in Figure 3.1c for the conditioning phase and in Figure 3.1d for the extinction phase. As predicted, my model tends to assign the conditioning and extinction trials to different clusters. When the test trial occurs in context A, the observation is assigned to the conditioning cluster; whereas when it occurs in context B, it is assigned to the extinction cluster. When the test trial occurs in a new context C, inference regarding its latent cause is divided between the conditioning and extinction clusters (and to a lesser extent a new cluster). This is due to the fact that as clusters accrue more observations, they come to dominate the posterior.

3.3.2 Latent inhibition

My model similarly explains the context-dependence of latent inhibition in terms of the partition structure of the animal's experience. I simulated latent inhibition with 15 pre-exposure trials and 15 conditioning trials (Figure 3.2a,b). When the animal receives pre-exposure ($\mathbf{f} = [\texttt{no reinforcement}, \texttt{tone}, \texttt{A}]$) and conditioning ($\mathbf{f} = [\texttt{reinforcement}, \texttt{tone}, \texttt{A}]$) in the same context, it is more likely to attribute a common latent cause to both phases, and thus the properties of both pre-exposure and conditioning observations are averaged together in making predictions about reinforcement in the conditioning phase, leading to a lower prediction and slower acquisition. In contrast, when the animal receives pre-exposure ($\mathbf{f} = [\texttt{no reinforcement}, \texttt{tone}, \texttt{A}]$) and conditioning ($\mathbf{f} = [\texttt{reinforcement}, \texttt{tone}, \texttt{B}]$) in different contexts, it is more likely to assign observations from each phase to different clusters—that is, to infer that different latent causes were active during pre-exposure and conditioning. In this case the reinforcement statistics learned from the conditioning trials are segregated from the reinforcement statistics of the pre-exposure

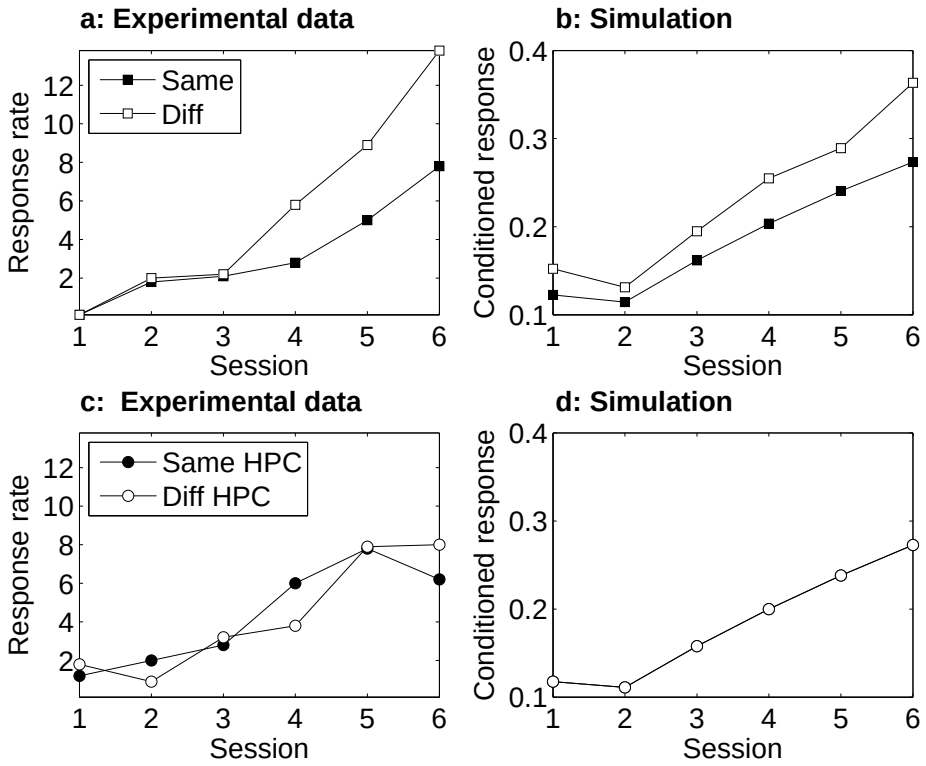

Figure 3.2: **Latent inhibition**. Experimental (*a,c*) and simulated (*b,d*) acquisition curves of conditioned responding to a stimulus paired with reinforcement as a function of whether unpaired stimulus pre-exposure occurred in the Same or in a different (Diff) context. (*a*) Pre-exposure in the same context as conditioning retards the acquisition of conditioned responding. This retarding effect is attenuated by pre-exposing the stimulus in a different context. (*b*) Simulated responding using the mixture model. (*c*) Subjects given hippocampal lesions before conditioning (HPC) show retarded acquisition regardless of whether pre-exposure is performed in the same or in a different context. Data replotted from Honey and Good (1993). (*d*) Simulated responding using the mixture model after hippocampal lesions, which were simulated by restricting the model's ability to infer new clusters. Note that Same HPC is indistinguishable from Diff HPC.

trials, eliminating the retarding effect of pre-exposure on learning, as can be seen in Figure 3.2a,b.

3.3.3 Pathologies of the model

Numerous studies have shown that damage to the hippocampus disrupts the context dependence of learning and extinction (for a review, see Ji and Maren, 2007). Animals

with pre-training electrolytic lesions of the dorsal hippocampus fail to show renewal of conditioned responding (Ji and Maren, 2005). Likewise, animals with hippocampal lesions exhibit intact latent inhibition even when pre-exposure and conditioning occur in different contexts (Honey and Good, 1993). These findings are paralleled by a similar lack of context-dependence in the behavior of the developing rat: before the age of ~ 22 days, rats do not show renewal or the attenuation of latent inhibition by conditioning in a new context (Yap and Richardson, 2005, 2007). I propose a unified explanation for these phenomena in terms of a pathology in my model's capacity to infer new latent causes. My theory additionally suggests an explanation for why the context-dependence of renewal and latent inhibition is only impaired when both the conditioning and extinction phases (for renewal) or pre-exposure and conditioning phases (for latent inhibition) occur before maturation (Yap and Richardson, 2005, 2007).

Hippocampal lesions

I propose that hippocampal lesions disrupt the ability of the animal to infer new clusters, restricting its inference to already-established clusters. I implemented this by setting α to zero at the time of the lesion. In latent inhibition, when this restriction was applied during pre-exposure, the pre-exposure and conditioning observations were assigned to the same cluster, regardless of the contexts that were in place for the two phases (in other words, my model degenerated into a single distribution over observation features). The prediction of reinforcement during conditioning was then based on an average of both pre-exposure and prior conditioning trials, leading to slower acquisition (Figure 3.2c,d).

Although early studies reported intact ABA renewal with pre-training electrolytic lesions of the fimbria/fornix (Wilson et al., 1995) or neurotoxic lesions of the entire hippocampus (Frohardt et al., 2000), Ji and Maren (2005) found that rats with pre-

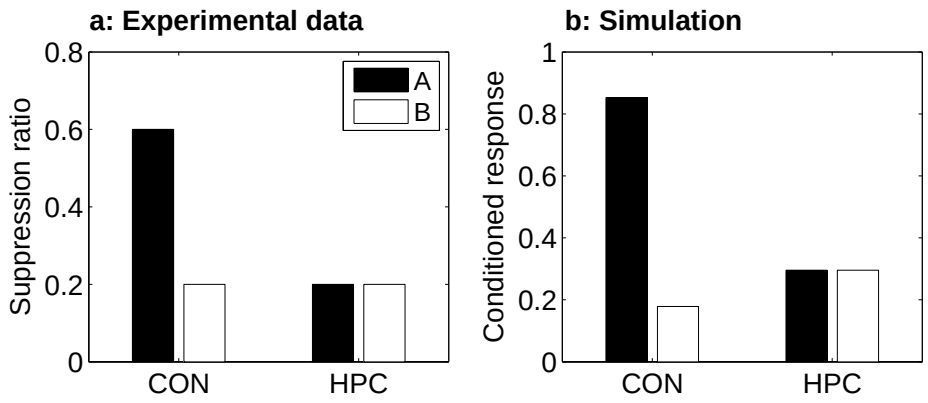

Figure 3.3: **Effect of hippocampal lesions on ABA renewal**. (*a*) Experimental conditioned responding to a cue during the test phase in control rats (CON) and those who received pre-training electrolytic lesions of the dorsal hippocampus (HPC). Data replotted from Ji and Maren (2005). (*b*) Simulated conditioned responding following restriction of the model's capacity to infer new clusters prior to training.

training electrolytic lesions of the dorsal hippocampus showed impaired renewal in the ABA paradigm.[7] Figure 3.3 shows these experimental data and simulated data from my model, demonstrating impaired renewal in my model after restricting the capacity to infer new clusters prior to training.

Developmental trajectories

Yap and Richardson (2005) have reported that in young rats latent inhibition is context independent, with behavior being strikingly similar to that exhibited by rats with pre-training hippocampal lesions. As shown in Figure 3.4a, when rats were pre-exposed, conditioned and tested at 18 days post-natal (PN18), they showed slow acquisition regardless of whether pre-exposure and conditioning occurred in the same or different contexts. In a second experiment, Yap and Richardson (2005) found that if testing was conducted at PN25, the context-independence of latent inhibition was still observed. In a third experiment, pre-exposure at PN18 with conditioning at PN24 and

[7]As discussed by Ji and Maren (2005), electrolytic lesions both damage neurons in the dorsal hippocampus and disrupt fibers of passage to subcortical structures, whereas fornix lesions only disrupt fibers of passage and neurotoxic lesions damage neurons while leaving fibers of passage intact. Thus, it is conceivable that these procedures failed to find an impairment in renewal because it is necessary to damage both fibers of passage and hippocampal neurons.

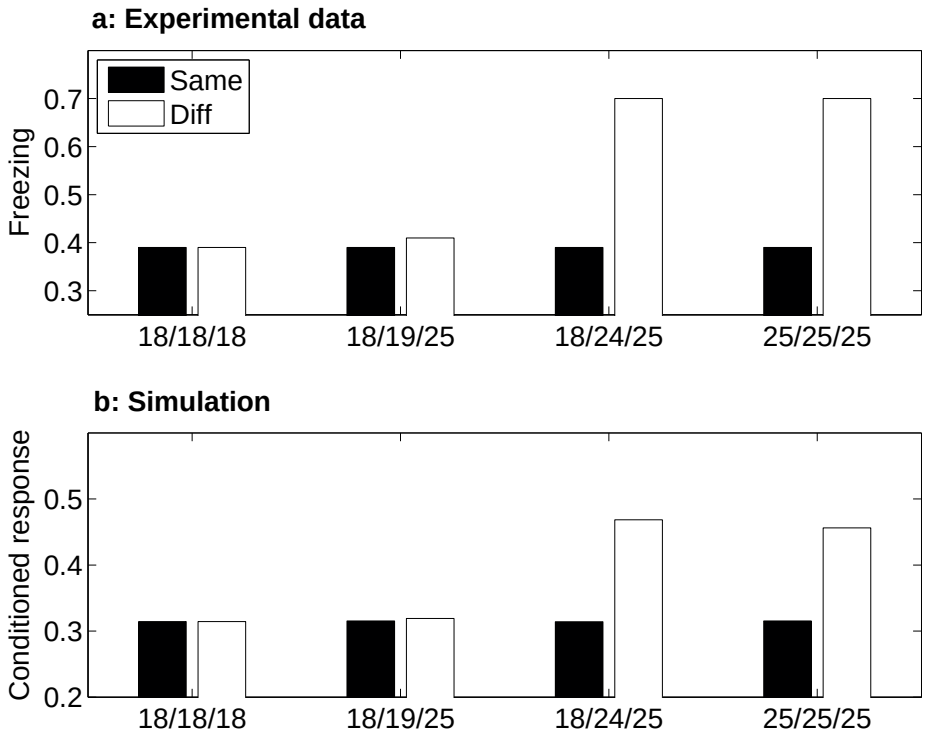

Figure 3.4: **Development of latent inhibition**. Experimental (a) and simulated (b) conditioned responding during the test phase following pre-exposure and conditioning in the same or in a different (Diff) context. Labels on the x-axis refer to the age at which each phase (pre-exposure/conditioning/test) was conducted. (a) Freezing to the stimulus in the test context. Data replotted from Yap and Richardson (2005). (b) Simulated conditioned responding.

testing at PN25 resulted in intact context-dependent latent inhibition. I simulated these different conditions by once again restricting my model's capacity to infer new clusters (setting $\alpha = 0$) during the phases when the animal is younger than PN22, and instating this capacity (setting $\alpha = 0.1$) when the animal reaches PN22. Figure 3.4b shows that with this manipulation the mixture model demonstrates a pattern of context-dependence similar to that observed experimentally. The explanation of these simulated results is the same as for the effects of pre-training hippocampal lesions described above.

Renewal has also been systematically investigated by Yap and Richardson (2007) in the developing rat. Figure 3.5a shows conditioned responding in contexts A and B

after conditioning in A and extinction in B at different ages, replotted from Yap and Richardson (2007). The main result is that if both conditioning and extinction are performed before maturity, no ABA renewal is observed, but if extinction is performed after maturity is reached, ABA renewal is intact. Figure 3.5b shows simulations of these experiments, demonstrating the same pattern of results. Only when my model's capacity for inferring new clusters is restricted during both conditioning and extinction will they be assigned to the same cluster. If extinction occurs after maturation, the animal can assign extinction observations to a new cluster, preventing interference between conditioning and extinction trials and thus enabling the conditioned response in the conditioning context to be renewed after extinction.

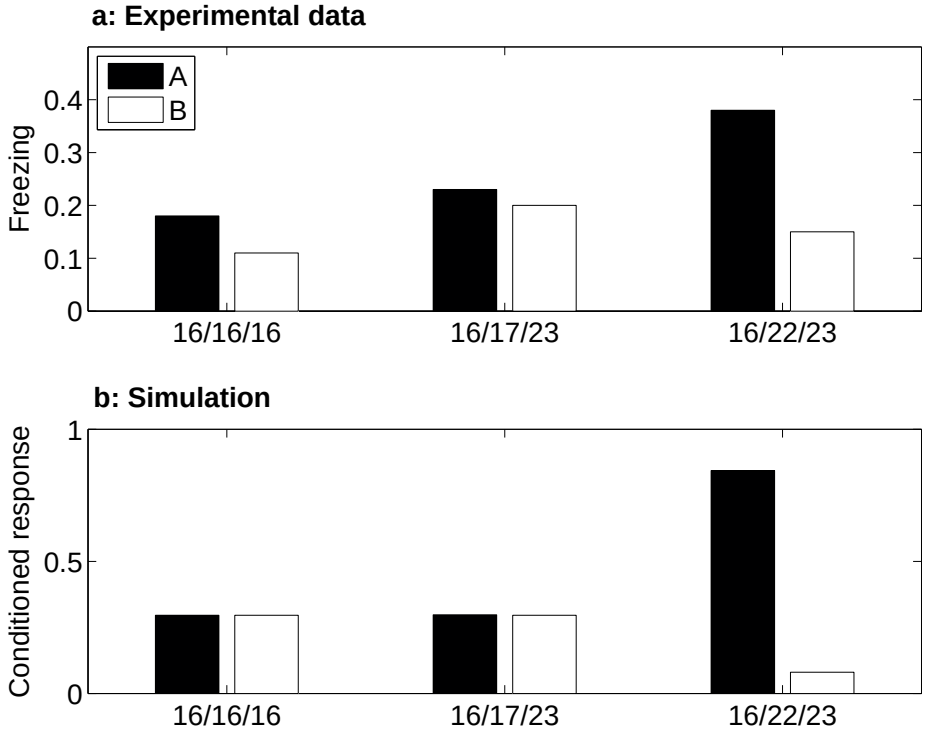

Figure 3.5: **Development of renewal**. Experimental (a) and simulated (b) conditioned responding during the test phase in context A or B following conditioning in A and extinction in B. Labels on the x-axis refer to the rat's age at each phase (conditioning/extinction/test). (a) Freezing to the cue in the test context. Data replotted from Yap and Richardson (2007). (b) Simulated conditioned responding.

3.4 Discussion

Starting from a normative statistical framework, I formalized a mixture model of animal learning in which context-dependent behavior is the result of inference over the latent causal structure of the environment. I showed that this model can explain several behavioral phenomena in latent inhibition and renewal paradigms. I also showed that restricting the model's capacity to infer new clusters can reproduce effects of hippocampal lesions and developmental changes in these paradigms.

My model extends and lends statistical clarification to the insights developed in the work of Redish et al. (2007). Additionally, I have addressed some specific shortcomings of their model. Importantly, the dependence of new state inference on negative prediction errors which prevented Redish et al.'s model from explaining the context specificity of latent inhibition is wholly absent from my account. I have also made a specific proposal regarding the role of the hippocampus in these tasks, which was alluded to in Redish et al. (2007), and will be discussed in more detail below. But first, to understand the theoretical motivation for a mixture model and its relationship to other models, it is useful to consider a taxonomy of models organized along four dimensions: computational problem, causal structure, capacity and inference algorithm. In the next four sections I detail each of these dimensions; I then discuss the role of hippocampus in learning and conclude with a discussion of limitations and possible extensions of my model.

3.4.1 Computational problem: generative vs. discriminative

Marr (1982) argued that to understand an information processing system, one must understand the computational problem it was designed to solve. Most models of animal learning are *discriminative*, implicitly or explicitly assuming that animals aim to predict the probability of reinforcement given the rest of their experience (Pearce

and Bouton, 2001). *Generative* models, in contrast, assume that animals aim to learn the joint distribution over all variables in their internal model of the environment, including, but not limited to, reinforcement. Courville (2006) has developed a generative model of animal learning using sigmoid belief nets that bears many similarities to my mixture model; I discuss similarities and differences in the following sections.

One might reasonably ask: why favor a generative account over a discriminative one? One problem with discriminative models is that they have no means of explaining behavioral phenomena in which the animal appears to learn information about the environment independently of reinforcement. A classic example of this is sensory preconditioning (Brogden, 1939): after initially pairing two neutral stimuli, A and B, in the absence of reinforcement, A is paired with reinforcement; subsequently, the animal exhibits significant responding to B, suggesting that an association between A and B was learned in the first phase despite the absence of reinforcement. Courville (2006) reviews numerous other phenomena that support a generative account.

Larrauri and Schmajuk (2008) have proposed a discriminative connectionist model to account for renewal and several other context-dependent behaviors. They argue that a combination of attentional, associative and configural mechanisms can collectively account for these data. As pointed out by Courville and colleagues (Courville et al., 2002; Courville, 2006), many of the ideas behind configural mechanisms can be captured by latent variable models. Whereas in connectionist models observation features are coupled via convergent projections to "configural" units, latent variable models capture this coupling generatively by having observation features share a common cause. Generalization between features is then accomplished by learning about their latent causes. In modeling the role of context in learning, I have adopted this same insight, showing how context can affect learning about reinforcers by means of a common latent cause.

A basic property of connectionist models such as that of Larrauri and Schmajuk

(2008) is that they effectively transpose the structure learning problem into a parameter learning problem by encoding all possible structures within the network, allowing the causal structure to be uncovered through experience-dependent adjustment of the connection weights (see also Gluck and Myers, 1993). One problem with this approach is that it ignores prior beliefs about the structure of the environment, which serve to constrain the kinds of structures that can be learned (Kemp and Tenenbaum, 2008; Courville et al., 2004). My generative model is a middle-ground between connectionist models that assume no prior structural beliefs and models that use hand-coded features for particular tasks (e.g., Brandon et al., 2000). In my model, where exactly in this middle-ground the animal's structural beliefs lie is determined both by its experience and its prior beliefs. The animal may initially expect only a small number of latent causes (specified by setting α close to 0), but its generative model is flexible enough to allow revision of this belief to accommodate more causes in light of new observations.

3.4.2 Causal structure: products vs. mixtures

Recall that in a mixture model observations are assumed to be caused by a generative model in which a single discrete cause is sampled and then an observation is sampled conditional on this cause. An alternative generative model, known as a *product* model, assumes that observations are generated by a linear combination of several latent causes, any number of which can be present at the same time.

As far back as the influential model of Rescorla and Wagner (1972), the predominant mathematical representation in models of animal learning is the product. In discriminative models, for example, the prediction of reinforcement is computed by taking a linear combination of feature variables. This is true not just for the Rescorla-Wagner model but for most models in the statistical and connectionist traditions as well (Dayan and Long, 1998; Schmajuk et al., 1996). In Courville's (2006) genera-

tive model, the probability of each observed variable is a linear combination of latent variables passed through a logistic sigmoid function.

An exception is the competitive mixture of experts model of Dayan and colleagues (Dayan and Long, 1998; Dayan et al., 2000), in which reinforcement is assumed to be generated by a single *observable* cause. In that model, the probability of reinforcement is the sum of conditioned probabilities of reinforcement given each cause, weighted by the probability of observing that cause (its "mixing probability"). One motivation for using mixtures rather than products, articulated by Dayan et al. (2000), is that inference within a mixture provides an elegant model of competitive attentional allocation in animal learning (whereby stimulus features are attended in proportion to the posterior probability that they caused reinforcement), and may be necessary to explain effects like downward unblocking (Holland, 1988). My model, while consistent with a competitive attentional account, puts mixtures to a different use by assuming that reinforcement is generated by a *latent* cause. There are many situations where this assumption is reasonable. Indeed, if one contemplates the designs of most conditioning experiments (including those modeled by Dayan and Long, 1998), the stimulus patterns presented to the animals are generated by discrete, latent phases of the experiment (e.g., conditioning, extinction); the animal never directly observes these phases, but inferring them is key to predicting reinforcement.

Fuhs and Touretzky (2007) have proposed a latent cause theory to explain hippocampal place cell remapping that is similar in spirit to my mixture model. They define context as a statistically stationary distribution of observations, and context learning as the task of clustering observations together into groups with local statistics that are stationary in time. In contrast to my static mixture model, they use a dynamic mixture model and formalize context learning in terms of Bayesian model selection, showing that this can predict when place cells will remap in response to environmental change. As with Courville's model, they select the best finite-capacity

mixture model (see next section), whereas I employ an infinite-capacity mixture model that automatically selects the number of clusters on the basis of its observations. Because they have applied this model to neural data and behavioral paradigms somewhat removed from my focus in this chapter, a direct comparison between the two models is difficult. Nonetheless, the idea that hippocampal place cells are important for inferring latent causes is consonant with the general view of the hippocampus set forth in this chapter.

3.4.3 Capacity: finite vs. infinite

A special problem vexes models with latent variables in which the number of latent variables is unknown. One can almost always increase the likelihood of the data under a model by increasing the number of parameters in the model. The number of parameters, or more generally the complexity of the model, is sometimes referred to as its "capacity." Increasing capacity can lead to a phenomenon known as "overfitting," wherein extra parameters are just capturing noise, leading to poor predictive power. A principled statistical approach to this problem is to represent uncertainty over the model's structure explicitly and infer both the structure and the values of the latent variables. This was the approach adopted by Courville (2006), who used an MCMC algorithm to select the best finite capacity model (a model with a fixed number of parameters) given the data.

An alternative to selecting between different finite capacity models is to allow the number of parameters to grow with the data (i.e., infinite capacity). This is, in fact, the spirit of Redish et al.'s (2007) model, in which heuristic modifications to a reinforcement learning algorithm allow it to increase its capacity (by expanding the state space) during learning. To control overfitting, one can place a prior distribution over parameters that expresses a preference for simpler models. This approach, adopted in my model, satisfies certain intuitions about an animal's representation of its en-

vironment. It seems unreasonable to assume that the animal knows in advance how many hidden causes it might be exposed to. A more reasonable assumption is that it infers that a new hidden cause is active when the statistics of its observations (e.g., lights, tones, odors) change, which is precisely the inference procedure imputed to the animal by the mixture model. Similar arguments have also been made by Sanborn et al. (2006) in their mixture model of human categorization.

Another aspect that my model shares with Redish et al.'s (2007) model is that cluster assignment (state classification) and cluster creation (expansion of the state space) are both determined by the similarity between the current observation and the existing states. A current observation is assigned to an existing cluster to the extent that it is similar to the other observations assigned to that cluster; if no cluster is sufficiently similar, a new cluster is created for that observation. In essence, the particle filter algorithm attempts to create clusters with maximal within-cluster similarity and minimal between-cluster similarity. The state classification mechanism in Redish et al.'s (2007) model also attempts to achieve this goal, but it lacks a direct statistical interpretation in terms of a well-defined inference procedure.

Redish et al.'s model does not represent uncertainty about the state classifications, whereas the particle filter maintains an approximation of the full posterior distribution over clusters.[8] This is potentially important in cases where previous clustering needs to be re-evaluated in light of later information. For example, imagine coming home and seeing the house flooded. You could classify this as either resulting from the (latent) cause "it has rained" or the *a priori* much less probable (latent) cause "there was a fire and fire trucks sprayed my house." Later hearing on the news that it had been an exceptionally hot and dry day, you might re-evaluate the fire hypothesis. Such *retrospective revaluation* phenomena (in which a previously disfavored interpretation becomes favored in light of new information) support the idea that humans and

[8]When the number of particles is small, particle filtering will behave similarly to hard assignment (see Daw and Courville, 2008; Sanborn et al., 2006).

animals represent uncertainty about past interpretations, rather than making hard assignments (Daw and Courville, 2008).

3.4.4 Inference algorithm: batch vs. incremental

One of the reasons for appealing to statistical models of learning is that they provide a formal description of the computational problem that the learning system is designed to solve. However, a complete analysis of an information processing system requires descriptions at two other levels (Marr, 1982). The *algorithmic* level specifies the operations and representations required to solve the computational problem. In a statistical model, the representations are probability distributions and the operations are usually some form of the product and sum rules from probability theory. The *implementational* level specifies how these computations are physically realized (e.g., in the brain). Statistical models of animal learning vary in their plausibility at these two levels of analysis (I discuss the implementational level in the next section).

At the algorithmic level, the main desideratum for plausibility is that the inference procedure be able to incorporate new data incrementally (Anderson, 1991; Sanborn et al., 2006). Reinforcement learning and connectionist updates satisfy this desideratum. The batch MCMC algorithm proposed by Courville et al. (2002, 2004), which must be re-run on all past observations after each trial, suffers in this regard, although later work attempted to remedy this drawback (Daw and Courville, 2008; Courville, 2006). I used the particle filter to perform inference in my model because it provides a cognitively plausible incremental algorithm for animal learning. However, it would be premature to commit to the particle filter as an algorithmic-level description of the conditioning data that I model, since given the large number of particles I use, this algorithm will make essentially identical behavioral predictions to any other algorithm that adequately approximates the posterior (e.g., MCMC sampling). With fewer samples, the particle filter approximates the posterior only crudely. It has been

argued that this might be the reason for certain kinds of resource limitations on behavior (Daw and Courville, 2008; Sanborn et al., 2006; Brown and Steyvers, 2009); it is an open question whether such resource limitations are evident in the renewal or latent inhibition data.

3.4.5 The hippocampus and context

The hippocampus has long been implicated in context learning, but theories have differed in their formal characterization of this role (Hirsh, 1974; O'Keefe and Nadel, 1978; Jarrard, 1993; Fuhs and Touretzky, 2007; Hasselmo and Eichenbaum, 2005). Here I have proposed one possible role for the hippocampus in inferring latent causes. I showed that restricting my model's ability to infer new clusters results in behavior qualitatively similar to that observed in rats with hippocampal lesions (see also Love and Gureckis, 2007, for a similar interpretation of human data). I believe that the hippocampus is suited for this role, with its ability to extract sparse codes from sensory inputs, which could support the learning of discrete latent causes. In particular, sparse projections from the dentate gyrus to CA3 are thought to be crucial for *pattern separation* (Marr, 1971), an operation that could serve to separate different observations (inputs) into distinct activation patterns in CA3. When a partial pattern (e.g., a stimulus and context) is presented, the missing part of the pattern (e.g., reinforcer) is activated by means of recurrent connections in CA3, which may function as an attractor network (McNaughton and Morris, 1987). These attractors may thus correspond to inferred clusters, with new attractors being formed when the input statistics change dramatically.

My model may also shed new light on a long-standing question about the hippocampus and memory in general (Marr, 1971; McNaughton and Morris, 1987): When a new observation is made, under what circumstances is a new trace encoded or an old trace retrieved? My model frames this as a choice between assigning an obser-

vation to an existing cluster or to a new cluster. O'Reilly and McClelland (1994) extensively analyzed a model of the hippocampus and argued that its anatomical and physiological properties might serve to minimize the trade-off between pattern separation (encoding) and pattern completion (retrieval). My model offers a normative motivation for how this trade-off should be balanced on the basis of the animal's observation statistics and prior beliefs, and future work should be directed at connecting it to the underlying neurophysiological mechanisms identified by O'Reilly and McClelland (1994), as well as the roles of theta oscillations and cholinergic input discussed by Hasselmo et al. (2002).

I would like to emphasize that the ostensibly "non-statistical" functions of the hippocampus like rapid conjunctive encoding (McClelland et al., 1995) are not incompatible with a statistical account. Most distinctions of this sort have identified statistical learning with extraction of the covariation structure of sensory inputs by neocortex (but see Gluck and Myers, 1993). In neural network models, this is implemented through gradual synaptic weight change. My model attempts to broaden this view of statistical learning to include the learning of discrete partition structure, a function that I argued fits with existing computational models of the hippocampus.

The fact that infant rats show a lack of context-dependence similar to rats with hippocampal damage (Yap and Richardson, 2005, 2007) suggests that the same causal inference mechanism may underlie both phenomena (Rudy, 1993; Martin and Berthoz, 2002), but more research on the behavioral consequences of hippocampal maturation is needed to test this idea. Other brain structures, notably the prefrontal cortex, also undergo maturation during this period, and it is unclear what specific contributions they may make to context-dependent learning and extinction (Quirk et al., 2006).

I have also said little about one of the main motivations for Redish et al.'s (2007) model, namely the role of the dopamine system in learning. Evidence has begun to accumulate suggesting that the hippocampal and dopamine systems are intricately

intertwined (Lisman and Grace, 2005); however, the behavioral significance of this relationship is poorly understood (but see Foster et al., 2000; Johnson et al., 2007). Finally, it is important to note that I do not view the role of the hippocampus in causal inference suggested here to be an all-encompassing functional description of the hippocampus. The hippocampus may perform several functions, or some more general function that includes causal inference as a sub-component. Furthermore, inference may rely on the interaction between the hippocampus and other regions in the medial temporal lobe and elsewhere (Corbit and Balleine, 2000).

3.4.6 Limitations and extensions

In their paper, Redish et al. (2007) also model the Partial Reinforcement Extinction Effect (PREE), the observation that extinction is slower when stimuli are only intermittently paired with reinforcement during training (Capaldi, 1957). My model, without further assumptions, cannot model this effect which depends crucially on using reinforcement *rate* as a contextual cue. My model assumes that reinforcements across trials are conditionally independent give their latent causes, and thus it has no representation of reinforcement rate. The essential explanation given by Redish et al. and others (e.g., Courville, 2006) is that the training and extinction contexts are harder to discriminate in the partial reinforcement condition due to smaller differences in reinforcement rate, and thus extinction trials are less likely to be assigned to a new cluster. Redish et al. (2007) were able to show this effect primarily because they included the time since last reinforcement, which is inversely correlated with reinforcement rate, in their prototype representation. I have found in simulations (not shown here) that augmenting the observation vector with an additional contextual feature that differs between training and extinction (which could be interpreted as a reinforcement rate cue) is sufficient to produce the PREE. However, an alternative approach to modeling this phenomenon is to incorporate an explicit model of dynamics

and change over time. Other extinction phenomena also depend on a richer representation of time than I have employed here. For example, in spontaneous recovery, simply waiting 48 hours after extinction is enough to produce renewed responding to the cue. The development of a temporally sophisticated mixture model is taken up in the next chapter.

Finally, I would like to note that although the formalism employed here appears to be a substantial departure from the type of reinforcement learning model used by Redish et al. (2007), the difference is not so great at it seems. Note that learning about reinforcement in my model essentially requires that the animal maintain and update a set of sufficient statistics about its beliefs—specifically, the average reinforcement in each cluster for each feature value. Such sufficient statistics might be learned by a mechanism similar to temporal difference learning (as I demonstrate in the next chapter), and hence may similarly rely on the dopamine system (see Daw et al., 2006, for related ideas).

3.5 Conclusions

I have argued that a wealth of behavioral data is consistent with an account of animal learning in which the animal infers the latent causes of its observations. Drawing on insights from Redish et al. (2007), I formalized this idea as a mixture model and showed how a particle filter algorithm can be used to perform inference. Simulations show that this framework can reproduce patterns of context-dependent behavior in latent inhibition and renewal paradigms. I also showed that restricting the model's ability to infer new clusters can reproduce patterns of hippocampal damage and developmental change. My model places context-dependent learning phenomena in a normative statistical framework, which I see as providing a computational-level analysis of the same problems addressed by Redish et al. (2007).

3.6 Appendix: particle filter algorithm

Recall that for trials $1 \dots t$ the vector $\mathbf{c}_{1:t}$ denotes a partition of the trials into clusters and $\mathbf{F}_{1:t}$ denotes the observations for these trials. The posterior approximation consists of m "particles," each corresponding to a hypothetical partition. In my implementation, the particles are generated by drawing m samples from the following distribution:

$$P(c_t^{(l)} = k) = \frac{1}{m} \sum_{l=1}^{m} P(c_t^{(l)} = k | \mathbf{c}_{1:t-1}^{(l)}, \mathbf{F}_{1:t}), \tag{3.3}$$

where $c_t^{(l)}$ denotes the latent cause for trial t in particle l, and

$$P(c_t^{(l)} = k | \mathbf{c}_{1:t-1}^{(l)}, \mathbf{F}_{1:t}) = \frac{P(c_t^{(l)} = k | \mathbf{c}_{1:t-1}^{(l)}) \prod_{i=1}^{D} P(f_{t,i} | c_t^{(l)} = k, \mathbf{c}_{1:t-1}^{(l)}, \mathbf{F}_{1:t-1})}{\sum_j P(c_t^{(l)} = j | \mathbf{c}_{1:t-1}^{(l)}) \prod_{i=1}^{D} P(f_{t,i} | c_t^{(l)} = j, \mathbf{c}_{1:t-1}^{(l)}, \mathbf{F}_{1:t-1})}. \tag{3.4}$$

The first term in Eq. 3.4 is the latent cause prior (Eq. 3.1). By default it is assumed that $c_1^{(l)} = 1$. The second term in Eq. 3.4 is the likelihood of the observed features on trial t given a hypothetical partition and the previous observations. Using a standard calculation for the Dirichlet-Multinomial model (Gelman et al., 2004), one can analytically integrate out the multinomial parameters ϕ associated with each cause to obtain the following expression for the likelihood:

$$\begin{aligned} P(f_{t,i} = j | c_t^{(l)} = k, \mathbf{c}_{1:t-1}^{(l)}, \mathbf{F}_{1:t-1}) &= \int_{\phi} P(f_{t,i} = j | c_t^{(l)} = k, \mathbf{c}_{1:t-1}^{(l)}, \mathbf{F}_{1:t-1}, \phi) P(\phi)\, d\phi \\ &= \frac{N_{i,j,k}^{(l)} + 1}{\sum_j (N_{i,j,k}^{(l)} + 1)}, \end{aligned} \tag{3.5}$$

where $N_{i,j,k}^{(l)}$ is the number of previous observations with value j on feature i that were generated by cause k in particle l (note that $N_{i,j,k}^{(l)}$ depends on $\mathbf{F}_{1:t-1}$).

The posterior over partitions is then approximated by an average of delta functions

placed at the particles:

$$P(\mathbf{c}_{1:t} = \mathbf{c}|\mathbf{F}_{1:t}) \approx \frac{1}{m}\sum_{l=1}^{m} \delta\left[\mathbf{c}_{1:t}^{(l)}, \mathbf{c}\right], \tag{3.6}$$

where $\delta[\cdot,\cdot]$ is 1 when its arguments are equal and 0 otherwise. As $m \to \infty$ this approximation converges to the true posterior. Although not immediately evident in these equations, learning occurs through maintaining and updating the sufficient statistics of each cluster, namely the cluster-feature co-occurrence counts (encoded by $N_{i,j,k}^{(l)}$).

Two things should be noted about this algorithm. First, hypothetical partitions are more likely to the extent that observations assigned to the same cluster are similar; this can be seen in Eq. 3.5. Second, the features interact multiplicatively in Eq. 3.4: a partition is more likely to the extent that *all* the observed features are likely under the particle's partition.

The probability of a US for a test observation (i.e., a feature vector in which the US feature is treated as missing data), which I denote by V_t, is calculated according to:

$$\begin{aligned} V_t &= P(f_{t,1} = \texttt{US}|\mathbf{f}_{t,2:D}, \mathbf{F}_{1:t-1}) \\ &= \sum_{\mathbf{c}_{1:t}} P(f_{t,1} = \texttt{US}|c_t, \mathbf{c}_{1:t-1}, \mathbf{f}_{1:t-1,1}) P(c_t|\mathbf{f}_{t,2:D}, \mathbf{F}_{1:t-1,2:D}, \mathbf{c}_{1:t-1}) P(\mathbf{c}_{1:t-1}|\mathbf{F}_{1:t-1}) \\ &\approx \frac{1}{m}\sum_{l=1}^{m}\sum_{k} r_{tk}^{(l)} P(f_{t,1} = \texttt{US}|c_t^{(l)} = k, \mathbf{c}_{1:t-1}^{(l)}, \mathbf{f}_{1:t-1,1}), \end{aligned} \tag{3.7}$$

where

$$r_{tk}^{(l)} = \frac{P(c_t^{(l)} = k|\mathbf{c}_{1:t-1}) \prod_{i=2}^{D} P(f_{t,i}|\mathbf{F}_{1:t-1}, \mathbf{c}_{1:t-1}^{(l)}, c_t^{(l)} = k)}{\sum_j P(c_t^{(l)} = j|\mathbf{c}_{1:t-1}) \prod_{i=2}^{D} P(f_{t,i}|\mathbf{F}_{1:t-1}, \mathbf{c}_{1:t-1}^{(l)}, c_t^{(l)} = j)}, \tag{3.8}$$

which is just Eq. 3.4 excluding the US feature in calculating the cluster assignment

probability.

Chapter 4

The computational nature of memory reconsolidation

When an experience is first written into memory, it is vulnerable to disruption by amnestic treatments or new learning, but over time the memory trace becomes progressively more resistant to disruption, a process known as "consolidation" (McGaugh, 2000; Muller and Pilzecker, 1900). This phenomenon raises a basic question about memory: once consolidated, can traces ever be modified again?

Answers to the contrary began to emerge several decades ago, beginning with a study demonstrating that retrieval of a memory can render it once again vulnerable to disruption, even after it has putatively consolidated (Misanin et al., 1968). Using a Pavlovian fear conditioning task, Misanin et al. (1968) found that electroconvulsive shock had no effect on fear memory administered one day after training; however, if the animal was briefly reexposed to the training cue prior to electroconvulsive shock, the animal subsequently exhibited loss of fear. This finding was followed by numerous similar demonstrations of what came to be known as *reconsolidation* (Spear, 1973), a term designed to emphasize the functional similarities between post-encoding and post-retrieval memory lability (see Riccio et al., 2006, for a historical overview).

Contemporary neuroscientific interest in reconsolidation was ignited by Nader et al. (2000), who showed that retrograde amnesia for an acquired fear memory could be produced by injection of a protein synthesis inhibitor (PSI) into the basolateral nucleus of the amygdala shortly after reexposure to the training cue. Subsequent studies have generated a detailed neural and behavioral characterization of reconsolidation, including a number of boundary conditions on the occurrence of reconsolidation (Nader and Hardt, 2009). Moreover, there is now evidence that retrograde amnesia can be obtained with a purely behavioral procedure (Monfils et al., 2009; Schiller et al., 2010). Despite these experimental advances, our understanding of reconsolidation remains largely descriptive in nature. As a consequence, many crucial mechanistic questions remain ambiguous or unanswered (Squire, 2006).

Reconsolidation challenges most existing models of Pavlovian conditioning. For concreteness, I focus on the most well-known of these, the Rescorla-Wagner model (Rescorla and Wagner, 1972), which posits that over the course of training the animal learns an association between the CS (e.g., tone) and the US (e.g., shock). The main weakness of the Rescorla-Wagner model is the assumption that presenting the stimulus repeatedly by itself (extinction) should erase the cue-outcome association formed during training—in other words, *extinction is unlearning*. It is now widely accepted that this assumption, shared by a large class of models, is wrong (Delamater, 2004).

As described in Chapter 1, Bouton (2004) reviewed a range of conditioning phenomena in which putatively extinguished associations are recovered. For example, simply increasing the time between extinction and test is sufficient to increase responding to the extinguished CS, a phenomenon known as *spontaneous recovery* (Pavlov, 1927; Rescorla, 2004). Another example is *reinstatement*: reexposure to the US alone prior to test increases conditioned responding to the CS (Pavlov, 1927; Rescorla and Heth, 1975; Bouton and Bolles, 1979b). Conditioned responding can

also be recovered if the animal is returned to the training context (Bouton and Bolles, 1979a).

Bouton (2004) interpreted the attenuation of responding after extinction in terms of a retrieval deficit that can be relieved by a change in temporal context or the presence of retrieval cues, thereby leading to recovery (see also Miller and Laborda, 2011). Central to retrieval-based accounts is the idea that the associations learned during training are largely unaffected by extinction because they are linked to the spatiotemporal context of the training session. Likewise, extinction results in learning that is linked to the spatiotemporal context of the extinction session. The manipulations reviewed above are hypothesized to either reinstate elements of the training context (e.g., renewal, reinstatement) or attenuate elements of the extinction context (e.g., spontaneous recovery).

Despite the qualitative appeal of this idea, no formal implementation has been shown to capture the full range of reconsolidation phenomena. The major stumbling block is that it is unclear what should constitute a spatiotemporal context: What are its constitutive elements, under what conditions are they invoked, and when should new elements come into play? In this chapter, I present a computational theory of Pavlovian conditioning that attempts to answer these questions, and use it to understand memory reconsolidation. I then show how this model can account for a wide variety of reconsolidation phenomena, including fine-grained temporal dynamics. This theory can be understood as a variant of the latent cause theory presented in Chapter 3.

Like the Rescorla-Wagner model (Figure 4.1A), my theory posits the learning of CS-US associations; however, these associations are modulated by the animal's beliefs about *latent causes*—hypothetical entities in the environment that interact with the CS and US (Courville, 2006; Courville et al., 2006; Gershman et al., 2010, 2012). I refer to the process of statistical inference over latent causes as *structure learning*,

whose interplay with associative learning determines the dynamics of reconsolidation. According to my theory, the animal learns a different set of associations for each cause, flexibly inferring new causes when existing causes no longer predict the CS-US relationship accurately (Figure 4.1B). This allows the theory to avoid the "extinction=unlearning" assumption by inferring that different latent causes are active during training and extinction, thus learning two sets of associations (see also Redish et al., 2007).

According to my theory, reconsolidation arises when CS reexposure provides evidence to the animal that the latent cause assigned to the training phase is once again active, making that cause's associations eligible for updating (or disruption by PSIs). I show that this theory is able to account for the main boundary conditions on reconsolidation using PSIs (Nader and Hardt, 2009), as well as the results of recent behavioral experiments (Monfils et al., 2009; Schiller et al., 2010). The theory also predicts a new boundary condition, which I confirm experimentally.

4.1 A rational analysis of Pavlovian conditioning

My theory is derived from a "rational analysis" (cf. Anderson, 1990) of the learning problem facing an animal in Pavlovian conditioning. To recapitulate the basic ideas introduced in Chapter 1, a rational analysis begins with a hypothetical generative process that describes how latent causes give rise to observed stimuli. The task of structure learning, according to my analysis, is to "invert" the generative process, using the observed stimuli to make inferences about the the latent causes that generated them (Gershman and Niv, 2010). The optimal inversion of the generative process is stipulated by Bayes' rule. The output of Bayes' rule is a posterior probability distribution over latent causes given the current sensory inputs. The posterior encodes the animal's belief about which cause generated its sensory inputs.

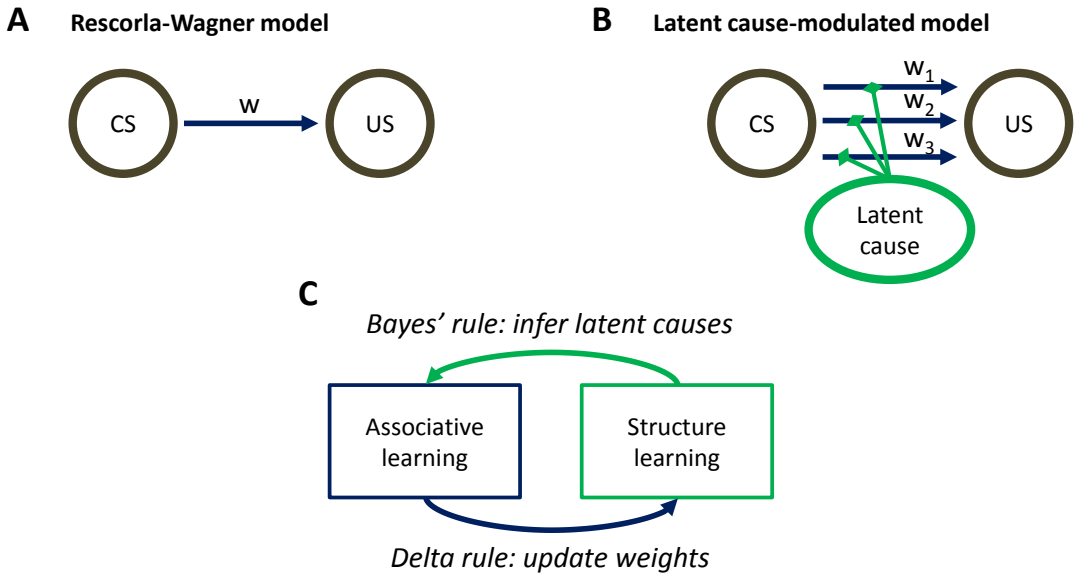

Figure 4.1: **Model schematic**. (A) The associative structure underlying the Rescorla-Wagner model. The associative strength between a conditional stimulus (CS) and an unconditional stimulus (US) is encoded by a scalar weight, w. (B) The associative structure underlying the latent cause-modulated model. As in the Rescorla-Wagner model, associative strength is encoded by a scalar weight, but in this case there is a collection of such weights, each paired with a different latent cause. The US prediction is a linear combination of weights, modulated by the posterior probability that the corresponding latent cause is active. (C) A high-level schematic of the computations in the latent cause model. Associative learning, in which the associative weights are updated (using the delta rule) conditional on the latent cause posterior, alternates with structure learning, in which the posterior is updated (using Bayes' rule) conditional on the weights.

4.1.1 High-level description of the theory

Before elaborating the technical details of my theory, I first provide a high-level description. The basic computational framework consists of two interacting sub-systems (Figure 4.1C): an associative learning system updates a set of CS-US associations using a delta rule (Widrow and Hoff, 1960; Rescorla and Wagner, 1972; Sutton and Barto, 1998), while a structure learning system updates an approximation of the posterior distribution over latent causes using Bayes' rule. It is useful to see the associative learning system as identical to the Rescorla-Wagner model, with the key difference that the system can maintain multiple sets of associations (one for each latent cause) instead of just a single set. Given a particular CS configuration (e.g., tone in a red box), the multiple associations are combined into a single prediction of the US by averaging the US prediction for each cause, weighted by the posterior probability of that cause being active. This posterior probability takes into account not only the conditional probability of the US given the CS configuration, but also the probability of observing the CS configuration itself. In the special case that only a single latent cause is inferred by the structure learning system, the associative learning system's computations are identical to the Rescorla-Wagner model (see the Appendix).

The structure learning system makes certain assumptions about the statistics of latent causes. Informally, the main assumptions I impute to the animal are summarized by two principles:

- *Simplicity principle*: sensory inputs tend to be generated by a small (but possibly unbounded) number of latent causes. The simplicity principle, or Occam's razor, has appeared throughout cognitive science in many forms (Chater and Vitányi, 2003). I use the CRP introduced in Chapter 2, an "infinite-capacity" prior over latent causes that, while preferring a small number of causes, allows the number of latent causes to grow as more data are observed.

- *Contiguity principle*: the closer two observations occur in time, the more likely they were generated by the same latent cause. In other words, *latent causes tend to persist in time.*

When combined with a number of additional (but less important) assumptions, these principles specify a complete generative distribution over sensory inputs and latent causes. I now describe the theory in greater technical detail.

4.1.2 The internal model

My specification of the animal's internal model consists of two parts: (1) a distribution over CS configurations, and (2) a conditional distribution over the US given the CS configuration. These two parts collectively define a joint distribution over the animal's sensory inputs.

I now introduce some notation to formalize these ideas. Let r_t denote the US at time t, and let $\mathbf{x}_t = [x_{t1}, \ldots, x_{tD}]$ denote the CS configuration. The distribution over r_t and $\mathbf{x}_t$ is determined by a latent cause vector $\mathbf{z}_t$, where $z_{tk} = 1$ if latent cause k is active on trial t and 0 otherwise. The latent cause vector is constrained so that only one latent cause is active on a given trial. I will sometimes abuse notation and use $z_t \in \{1, \ldots, K\}$ to denote the latent cause on trial t, where K denotes the maximal number of latent causes (as described below, this number can grow with new data).

Formally, the CS configuration is drawn from a Gaussian distribution:

$$P(\mathbf{x}_t | z_t = k) = \prod_{d=1}^{D} \mathcal{N}(x_{td}; \mu_{kd}, \sigma_x^2), \tag{4.1}$$

where μ_{kd} is the expected intensity of the dth CS given cause k is active, and σ_x^2 is its variance. I assume a zero-mean prior on μ_{kd} with a variance of 1, and treat σ_x^2 as a fixed parameter (see the Appendix). Similarly to the Kalman filter model of conditioning (Kakade and Dayan, 2002; Kruschke, 2008), I assume that the US is

generated by a weighted combination of the CSs corrupted by Gaussian noise:

$$r_t = \sum_k z_{tk} \sum_{d=1}^{D} w_{kd} x_{td} + \epsilon_t = \mathbf{z}_t \mathbf{W} \mathbf{x}_t^\top + \epsilon_t, \tag{4.2}$$

where $\epsilon_t \sim \mathcal{N}(0, \sigma_r^2)$ is a noise term.

I assume that on each trial a latent cause vector $\mathbf{z}_t$ is drawn from the following distribution:

$$P(z_{tk} = 1 | \mathbf{z}_{1:t-1}) \propto \begin{cases} \sum_{\tau<t} \mathcal{K}(t-\tau) & \text{if } k \leq K \text{ (i.e., } k \text{ is an old cause)} \\ \alpha & \text{otherwise (i.e., } k \text{ is a new cause)} \end{cases} \tag{4.3}$$

where τ ranges over the timepoints prior to t and $\mathcal{K}$ is a temporal kernel that governs the temporal dependence between latent causes. Intuitively, the CS configuration on a particular trial will be generated by the same latent cause as other trials that occurred nearby in time. The "concentration" parameter α determines the prior bias towards generating a new latent cause. This prior imposes the simplicity principle described in the previous section—a small number of latent causes is favored *a priori* over a large number. The distribution defined by Eq. 4.3 was first introduced by Zhu et al. (2005) in their "time-sensitive" generalization of the CRP (Aldous, 1985).[1] Gershman and Blei (2012) for a tutorial introduction.

I use a power law kernel, $\mathcal{K}(t-\tau) = (t-\tau)^{-1}$, which has an important temporal compression property (illustrated in Figure 4.2). Consider two timepoints, $t_1 < t_2$, separated by a fixed temporal distance, $t_2 - t_1$, and a third time point, $t_3 > t_2$, separated from t_2 by a variable interval, $t_3 - t_2$. In general, the same latent cause is more likely to have generated both t_2 and t_3 than t_1 and t_3 (the contiguity principle). However, this advantage diminishes over time, and asymptotically disappears: As $t_3 - t_2$ is increased, holding $t_2 - t_1$ constant, the ratio $P(z_3 = z_2)/P(z_3 = z_1)$ decreases

[1] It is also equivalent to a special case of the "distance dependent" CRP described by Blei and Frazier (2011).

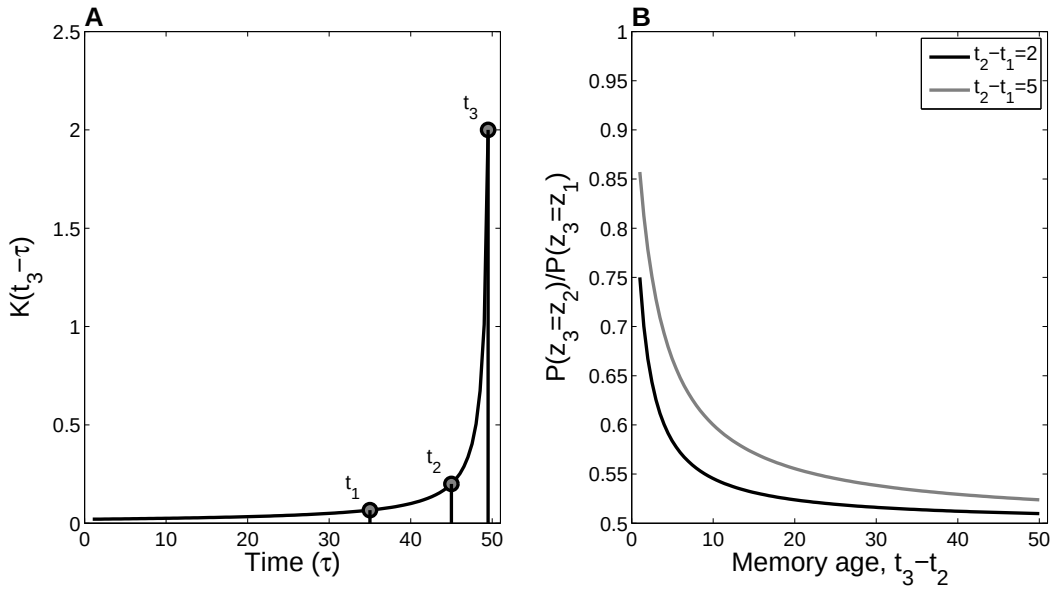

Figure 4.2: **Temporal compression with the power law kernel**. (A) The power law temporal kernel $\mathcal{K}(t-\tau) = (t-\tau)^{-1}$. Three timepoints (t_1, t_2, t_3) are shown for illustration. (B) As timepoints t_1 and t_2 (separated by a fixed temporal difference) recede into the past relative to time point t_3, the probability of the same latent cause generating t_3 and t_2 diminishes.

monotonically to 0.5 according to Eq. 4.3.

This completes my description of the animal's internal model. In the next section, I describe how an animal can use this internal model to reason about the latent causes of its sensory inputs and adjust the model parameters to improve its predictions.

4.1.3 Associative and Structure Learning

According to my rational analysis, two computational problems confront the animal: (1) associative learning, the adjustment of the model parameters (specifically, the associative weights, $\mathbf{W}$) to maximize the likelihood of the observations under the generative model, and (2) structure learning, the assignment of observations to latent causes. The alternation of these two learning processes can be understood as a variant of the expectation-maximization (EM) algorithm (Dempster et al., 1977; Neal and Hinton, 1998). Friston (2005) has argued that the EM algorithm provides a unifying

framework for understanding cortical computation.

For my model, the EM algorithm takes the following form (see Appendix for details). After each observation, the model alternates between structure learning (the E-step, in which the posterior distribution over latent causes is updated) and associative learning (the M-step, in which the weights are updated):

$$\textbf{E-step} : q_{tk}^{n+1} = P(z_{tk} = 1|\mathcal{D}_{1:t}, \mathbf{W}^n)$$
$$\textbf{M-step} : w_{kd}^{n+1} = w_{kd}^n + \eta x_{td}\delta_{tk}$$

where η is a learning rate and

$$\delta_{tk} = q_{tk}^{n+1}(r_t - \textstyle\sum_d w_{kd}x_{td}) \tag{4.4}$$

is the prediction error at time t.

Associative learning (the M-step of the EM algorithm) in my model is a generalization of the Rescorla-Wagner model (see the Appendix for further details). Whereas in the Rescorla-Wagner model there is a single association between a CS and the US (Figure 1A), in my generalization the animal can form multiple associations depending on the latent causes it infers (Figure 4.1B). The optimal US prediction is then a weighted combination of the CSs, where the weights are modulated by the posterior probability distribution over latent causes, represented by q. Associative learning proceeds by adjusting the weights using gradient descent to minimize the prediction error.

Structure learning (the E-step of the EM algorithm) consists of computing the posterior probability distribution over latent causes using Bayes' rule:

$$P(z_t = k|\mathcal{D}_{1:t}, \mathbf{W}^n) = \frac{P(\mathcal{D}_{1:t}|z_t = k, \mathbf{W}^n)P(z_t = k)}{\sum_j P(\mathcal{D}_{1:t}|z_t = j, \mathbf{W}^n)P(z_t = j)}. \tag{4.5}$$

The first term on the right-hand side of the numerator is the *likelihood*, encoding the probability of the animal's observations are under a hypothetical latent cause assignment, and the second term is the *prior* (Eq. 4.3), encoding the animal's inductive bias about which latent causes are likely to be active. As explained in the Appendix, Bayes' rule is in this case computationally intractable (due to the implicit summation over the history of previous latent cause assignments, $\mathbf{z}_{1:t-1}$); I therefore use a simple and effective approximation (see Eq. 4.11).

Because the E and M steps are coupled, they need to be alternated until convergence (Figure 4.1C). Intuitively, this corresponds to a kind of offline "rumination," in which the animal continues to revise its beliefs even after the stimulus has disappeared. In the context of Pavlovian conditioning, I assume that this happens during intervals between trials, up to some maximum number of iterations (i.e., until the animal starts thinking about something else). In my simulations, I take this maximum number to be 3, which corresponds roughly to a few minutes in the timescale adopted by my simulations.[2] The explanatory role of multiple iterations comes into play when I discuss the Monfils-Schiller paradigm below.

4.1.4 Prediction

The animal's prediction of the US on trial t is given by:

$$\tilde{r}_t = \mathbb{E}[r_t|\mathbf{x}_t, \mathcal{D}_{1:t-1})] = \sum_{d=1}^{D} x_{td} \sum_k w_{kd} P(z_t = k|\mathbf{x}_t, \mathcal{D}_{1:t-1}). \tag{4.6}$$

Note that the posterior probability in this equation does not condition on r_t, whereas the posterior used for structure learning (Eq. 4.5) *does* condition on r_t. Most earlier Bayesian models of conditioning assumed that the animal's conditioned response is directly proportional to the expected reward (e.g., Courville, 2006; Gershman and

[2]While the qualitative structure of the theory's predictions does not depend strongly on this maximum number, I found this to produce the best match with empirical data.

Niv, 2010; Kakade and Dayan, 2002). In my simulations, I found that while Eq. 4.6 generally agrees with the direction of empirically observed behavior, the predicted magnitude of these effects was not always accurate. One possible reason for this is that in fear conditioning the mapping from predicted outcome to behavioral response may be nonlinear (e.g., an all-or-none response). I therefore use a nonlinear sigmoidal transformation of Eq. 4.6 to model the conditioned response:

$$\mathrm{CR} = 1 - \Phi(\theta; \tilde{r}_t, \lambda), \tag{4.7}$$

where $\Phi(\cdot)$ is the Gaussian cumulative distribution function. One way to understand Eq. 4.7 is that the animal will emit a conditioned response if the predicted US is greater than some threshold, θ. When $\lambda = \sigma_r^2$, Eq. 4.7 corresponds precisely to the posterior probability that the US exceeds θ:

$$\mathrm{CR} = P(r_t > \theta | \mathbf{x}_t, \mathcal{D}_{1:t}) = \int_\theta^\infty P(r_t | \mathbf{x}_t, \mathcal{D}_{1:t}) dr_t. \tag{4.8}$$

In practice, I found that descriptively more accurate results could be obtained by setting $\lambda < \sigma_r^2$. At a mechanistic level, λ functions as an inverse gain control parameter: larger values of λ generate more sharply nonlinear responses (approaching a step function as $\lambda \to 0$).

4.2 Understanding Extinction and Recovery

Before modeling specific experimental paradigms, in this section I lay out some general intuitions for how my model deals with extinction and recovery. In previous work (Gershman et al., 2010), I argued that the transition from training to extinction involves a dramatic change in the statistics of the animal's sensory inputs, leading the animal to assign different latent causes to training and extinction. The result of this

partitioning is that the training associations are not unlearned during extinction, and hence can be later recovered, as is observed experimentally (Bouton, 2004). Thus, according to my model, the key to enduring extinction (i.e., erasure of the CS-US association learned during training) is to finesse the animal's observation statistics such that the posterior favors assigning the same latent cause to both training and extinction phases.

One way to understand the factors influencing the posterior is in terms of prediction error, the discrepancy between what the animal expects and what it observes. This typically refers to a US prediction error, but my analysis applies to CS prediction errors as well. The prediction error plays two roles in my model: as an associative learning signal that teaches the animal how to adjust its associative weights, and as a segmentation signal indicating when a new latent cause is active. When the animal has experienced several CS-US pairs during training, it develops an expectation that is then violated during extinction, producing a prediction error. This prediction error can be reduced in two different ways: either by associative learning (unlearning the CS-US association) or by structure learning (assigning the extinction trials to a new latent cause). Initially, the prior simplicity bias towards a small number of latent causes favors unlearning, but a persistent accumulation of these prediction errors over the course of extinction eventually makes the posterior probability of a new cause high. Thus, standard training and extinction procedures lead to the formation of two memories, one for CS-US and one for CS-noUS.

The opposing effects of prediction errors on associative and structure learning are illustrated in Figure 4.3. If the prediction errors are too small, the posterior probability of the training latent cause will be high (leading to memory modification) but the amount of CS-US weight change will be small; if the prediction errors are too big, the posterior probability of the training latent cause will be low (leading to memory formation), and the change in the corresponding weight will again be small.

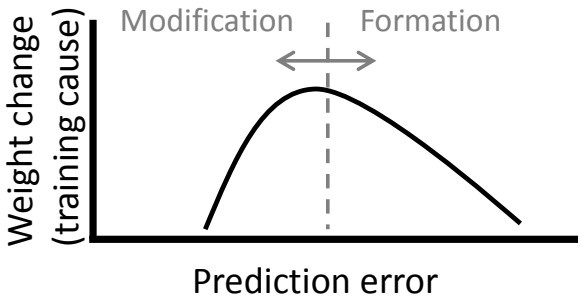

Figure 4.3: Cartoon of the model's predictions for fear extinction. The X-axis represents the size of the prediction error (experienced minus expected reward) during extinction, and the Y-axis represents the change (after learning) in the weight corresponding to the "training latent cause" (i.e., the latent cause inferred by the animal during fear conditioning).

There exists an intermediate "sweet spot" where the prediction errors are large enough to induce weight change but small enough to avoid inferring a new latent cause. In the next section, I describe an experimental paradigm that, according to my theory, achieves this sweet spot.

To get a feeling for how the model's response to prediction errors depends on the parameter settings, I can solve explicitly for the "prediction error threshold"—the value of the squared prediction error[3] at which a new latent cause will be inferred. For simplicity, let us assume that a single cue has been paired N times with reward ($\mathcal{D}_{1:N} = \{x_t = 1, r_t = 1\}_{t=1}^{N}$). Under most parameter settings, this will result in all the training trials being assigned to a single latent cause (hence I ignore the cause subscript k in this example). Now consider what happens when a single extinction trial ($x_{N+1} = 1, r_{N+1} = 0$) is presented. Using the posterior approximation described in the Appendix, if the squared prediction error δ_{N+1}^2 is greater than a certain threshold $\tilde{\delta}_{N+1}^2$, the extinction trial will be assigned to a new latent cause. Holding the

[3]I analyze the squared prediction error because I am concerned with magnitude rather than sign here.

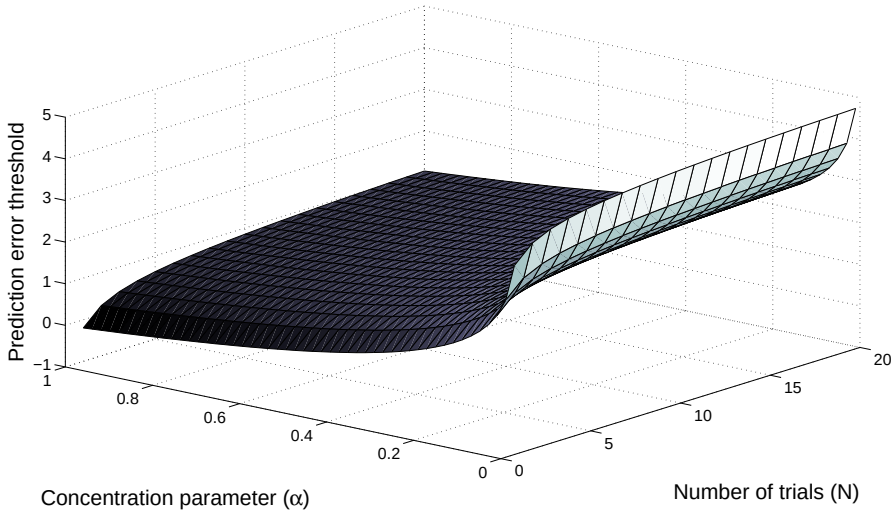

Figure 4.4: **Prediction error threshold**. When the squared prediction error exceeds this threshold (within the simplified example described in the text), a new latent cause is inferred. For this plot, $\sigma_r^2 = 0.4$ (the same values as used in all the simulations).

other parameters fixed, this threshold can be solved algebraically, giving:

$$\tilde{\delta}^2_{N+1} = 2\sigma_r^2 \log \left[\frac{\mathcal{N}(1; 0, \sigma_x^2)}{\mathcal{N}(1; \hat{x}_N, \nu_N^2)} \frac{\sum_{t=1}^{N} \mathcal{K}(t)}{\alpha \sigma_r \sqrt{2\pi}} \right]. \tag{4.9}$$

This threshold is plotted as a function of the concentration parameter α and N in Figure 4.4. As the concentration parameter is increased, the prior bias for simplicity (a small number of latent causes) decreases, making it more likely that the prediction error was due to a new latent cause being active; this results in a lower prediction error threshold. As N increases, the internal model becomes more confident that only a single latent cause is active (due to the homogeneity of the sensory statistics), resulting in an increasing threshold as a function of N.

In order to understand some of the empirical phenomena described below, I must also explain why spontaneous recovery occurs in my model: Why does the posterior probability of the training cause increase as the extinction-test (aka retention) interval is lengthened? The answer lies in my choice of temporal kernel $\mathcal{K}(t)$ as a power law, which (as explained above) has the important property that older timepoints are

"compressed" together in memory: latent causes become more equiprobable under the prior as the time between training and test increases.[4] Thus, the prior advantage of the extinction cause over the training cause diminishes with the retention interval. One implication of this analysis is that spontaneous recovery should never be complete, since the prior probability of the training cause can never *exceed* the probability of the extinction cause (though the ratio of probabilities increases monotonically towards 1 with the retention interval); this appears generally consistent with empirical data (Rescorla, 2004).

Another important feature of my model is that both associative and structure learning can occur "offline" (i.e., between trials). This has several empirical implications for extinction. First, a memory may be incrementally unlearned during the intertrial interval, due to repeated iterations of the M-step. A stronger memory will require more iterations to incrementally unlearn. Second, a trial may be initially assigned to a new latent cause (i.e., on the E-step of the first iteration), but this assignment may change as adjustments to the initial memory are made in the M-step. In particular, the M-step can alter the acquisition memory so as to make it more similar to the extinction trials by unlearning the CS-US association. I will return to this idea in my simulations.

4.3 Boundary Conditions on Reconsolidation

In this section, I explore several boundary conditions on reconsolidation (see Nader and Hardt, 2009, for a review). My goal is to show that these conditions fall naturally out of my rational treatment of Pavlovian conditioning. I seek to capture the *qualitative* pattern of results, rather than their precise quantitative form. I thus use the same parameters for all simulations, rather than fitting the parameters to data.

[4] A similar idea was used by Brown et al. (2007) in their model of episodic memory to explain recency effects in human list learning experiments.

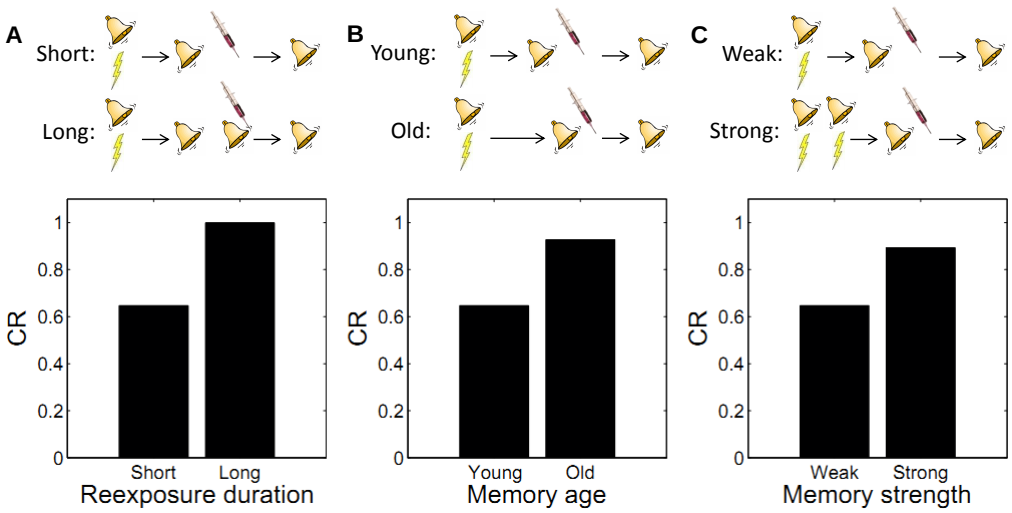

Figure 4.5: **Boundary conditions on reconsolidation**. Top row shows a schematic of the experimental design (bell represents the tone CS, lightning bolt represents the shock US, syringe represents the injection of a protein synthesis inhibitor). Bottom row shows model predictions in the test phase. Memory updating is attenuated under conditions of (A) longer reexposure, (B) older or (C) stronger memories.

Many of the experiments used PSIs administered shortly after cue reexposure as the amnestic agent; accordingly, I modeled PSI injections by decrementing the weights according to: $\mathbf{w}_k \leftarrow \mathbf{w}_k(1 - q_{tk})$. In other words, I decremented the weights for latent cause k towards 0 in proportion to the posterior probability that cause k is active on trial t. This is essentially a formalization of the *trace dominance principle* proposed by Eisenberg et al. (2003): memories will be more affected by amnestic agents to the extent that they control behavior at the time of treatment (see below).

The trace dominance principle. Using fear conditioning in the Medaka fish, Eisenberg et al. (2003) found that applying an amnestic agent after a single re-exposure to the CS (i.e., a single extinction trial) caused retrograde amnesia for the reactivated fear memory, but applying the amnestic agent after multiple re-exposures caused retrograde amnesia for extinction (i.e., spontaneous recovery is observed after 24 hours). Similar results have been obtained with mice (Suzuki et al., 2004), rats (Lee et al., 2006), and the crab *Chasmagnathus* (Pedreira and Maldonado, 2003). This pattern

of results is consistent with the trace dominance principle, under the assumption that reexposure duration determines the dominance of a memory. In terms of my model, a short reexposure duration favors an assignment of the reexposure trial to the training latent cause. This follows from the simplicity bias in the latent cause prior: In the absence of strong evidence to the contrary, the prior prefers assigning new observations to previously inferred causes. However, with longer durations, the evidence favoring a new latent cause (accruing from persistent prediction errors) overwhelms the prior, favoring assignment to a new latent cause. This logic leads to model predictions consistent with the empirical data (Figure 4.5A).

Memory age. By manipulating the interval between training and reexposure, Suzuki et al. (2004) demonstrated that the amnestic effects of PSI injection were more pronounced for short retention intervals (i.e., young memories). Winters et al. (2009) found a similar effect with the NMDA receptor antagonist MK-801 administered prior to re-exposure, and Milekic and Alberini (2002) demonstrated this effect in an inhibitory avoidance paradigm. Alberini (2007) has reviewed several other lines of evidence for the age-dependence of reconsolidation. These findings can be explained by my model: old observations are less likely to have been generated by the same latent cause as recent observations under the prior. Thus, there is an inductive bias against modifying old memory traces. Figure 4.5B shows simulations of the Suzuki paradigm, demonstrating that my model can reproduce this pattern of results.

Memory strength. In another experiment, Suzuki et al. (2004) showed that strong memories are more resistant to updating (see also Wang et al., 2009). Specifically, increasing the number of training trials led to persistent fear even after PSI injection. In terms of my model, this phenomenon reflects the fact that for stronger memories, it takes more iterations to incrementally reduce the CS-US association to a low level. Consequently, the interplay between associative and structure learning described in the previous section will tend to favor inferring a new cause. Simulations of this

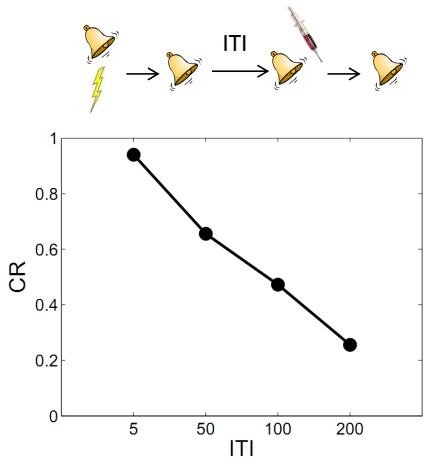

Figure 4.6: **Timing of multiple reexposures**. Lengthening the intertrial interval (ITI) between multiple reexposures increases the effectiveness of PSI administration.

experiment (Figure 4.5C) demonstrate that stronger memories are more resistant to updating in my model.

Timing of multiple reexposures. When two CSs are reexposed with a short ITI separating them, PSI injection following the second CS fails to disrupt reconsolidation (Jarome et al., 2012). This is essentially another manifestation of the trace dominance principle (Eisenberg et al., 2003): two unreinforced reexposures cause the extinction trace to become dominant, and the PSI therefore disrupts the extinction trace rather than the fear trace. Jarome et al. (2012) found that increasing the ITI results in a parametric decrease of fear at test, suggesting that longer intervals lead to disruption of the fear trace by the PSI. This effect is predicted by my theory, because longer ITIs reduce the probability that the two reexposures were generated by the same "extinction" latent cause, concomitantly increasing the probability that the second reexposure was generated by the "training" latent cause (Figure 4.6). The explanatory work here is being done by the time-dependent prior over latent causes, which prefers assigning trials separated by a long temporal interval to different causes.

Prediction Error and Novelty. As described above, prediction errors play two central roles in my model, driving both associative and structure learning. Of particular

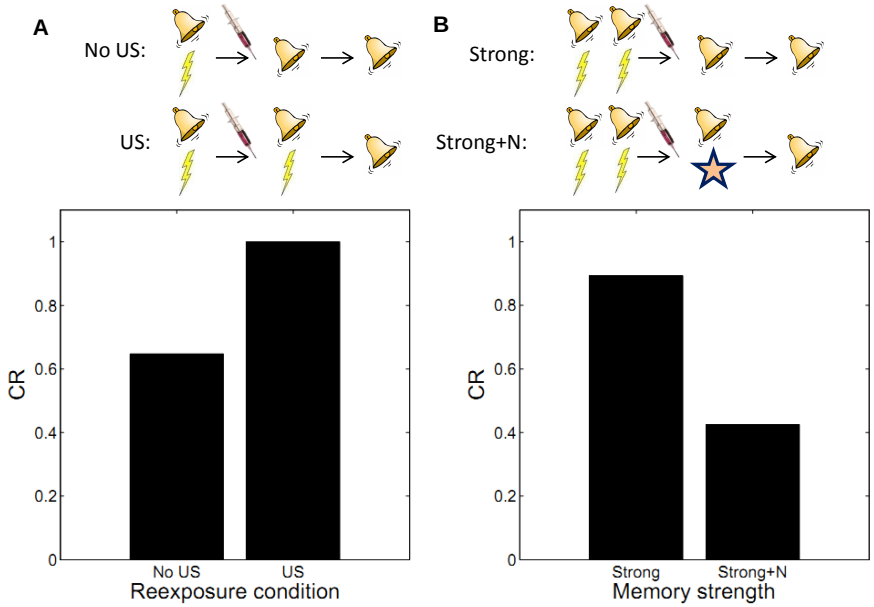

Figure 4.7: **The role of prediction error in reconsolidation**. The role of prediction error in reconsolidation. (A) Presenting the US during reexposure prevents reconsolidation. (B) Strong memories can be reconsolidated when reexposure is accompanied by a novel object (Strong+N, indicated by a star), thereby eliminating the strength-based boundary condition.

relevance to this claim is research showing that violation of the animal's expectations (i.e., prediction error) is necessary to induce memory updating (Pedreira et al., 2004; Morris et al., 2006; Winters et al., 2009). In one manifestation of this boundary condition, Pedreira et al. (2004) found that updating does not occur when the retrieval trial is reinforced. This finding is consistent with my model (Figure 4.7A), which predicts that a new latent cause will only be inferred when there is some prediction error. This prediction error can also be induced by introducing novel stimuli (Morris et al., 2006; Winters et al., 2009). For example, Winters et al. (2009) showed that adding a novel object can eliminate the memory strength boundary condition: strong object memories can be updated after training if the object is paired with novelty. I simulated this by adding a novel CS during the retrieval trial; Figure 4.7B shows that a strong memory is sensitive to disruption when accompanied by novelty.

Cue-specificity. Doyère et al. (2007) reported that disruption of reconsolidation

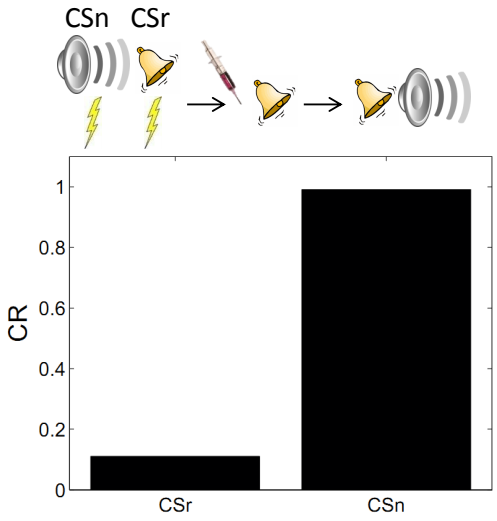

Figure 4.8: **Cue-specificity of amnestic treatment**. Disruption of reconsolidation by amnestic treatment affects the reactivated cue (CSr) but not the non-reactivated cue (CSn).

by an amnestic treatment (in this case the mitogen-activated protein kinase inhibitor U0126) is restricted to a reactivated CS, leaving intact the CR to a non-reactivated CS that had also been paired with the US (Figure 4.8). This effect arises in the model because learning only occurs for associations specific to the current CS and inferred latent cause.

Transience of amnesia. A major focus of retrieval-based theories of reconsolidation has been the observation that, under a variety of circumstances, recovery from amnesia can be observed (Miller and Matzel, 2006; Riccio et al., 2006). Within the recent wave of research, a study by Power et al. (2006) provides a clear demonstration: Following inhibitory avoidance training, intrahippocampal infusions of the PSI anisomycin impaired memory retention when the rats were tested 1 day later, but memory was unimpaired when the test was administered after 6 days. Thus, the PSI-induced memory impairment was transient (see also Lattal and Abel, 2004). As pointed out by Gold and King (1974), recovery from amnesia does not necessarily mean that the amnesia was purely a retrieval deficit. If the amnestic agent

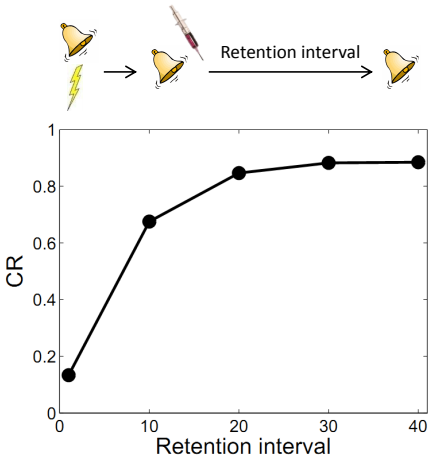

Figure 4.9: **Transience of amnesia**. Lengthening the retention interval between reexposure and test produces recovery from amnesia.

diminished, but did not entirely eliminate, the reactivated memory, then subsequent recovery could reflect new learning added onto the residual memory trace[5]. The explanation offered by my theory is related: Since the amnestic agent does not entirely eliminate the memory trace, recovery occurs because the relative probability of assigning a new observation to the training cause increases over time (a consequence of temporal compression by the power law kernel, as explained above). Simulations shown in Figure 4.9 demonstrate that this explanation can account for the increase in CR with longer retention intervals.

4.4 The Monfils-Schiller Paradigm

In two recent studies (Monfils et al., 2009; Schiller et al., 2010), it was demonstrated that a single CS presentation ("retrieval trial") 10-60 minutes before extinction leads to apparent memory erasure (as measured by renewal, reinstatement and spontaneous recovery tests). These studies also revealed several other effects of this paradigm: (1)

[5]This assumes that nonreinforced presentations of the CS can evoke a memory of past reinforcements, thereby paradoxically strengthening the memory (see Eysenck, 1968; Rohrbaugh and Riccio, 1970).

extinction of fear lasts up to a year later; (2) erasure is specific to the reactivated memory; and (3) increasing the retrieval-extinction interval to 6 hours eliminates the effect. This latter finding suggests that a time-limited plasticity window is engaged by the retrieval trial. These findings have been recently replicated in rats (Clem and Huganir, 2010) and humans (Oyarzún et al., 2012), though the generality of these findings remains controversial (Chan et al., 2010; Costanzi et al., 2011; Kindt and Soeter, 2011).

It is important to recognize that there are only two salient differences between the Monfils-Schiller paradigm and regular extinction training: (1) The lengthened interval between the 1st and 2nd extinction trials; and (2) the subject spent this interval outside the training context. My theoretical explanation of these data thus rests critically on what happens during the interval between the 1st and 2nd extinction trials. This phenomenon is puzzling for most—if not all—theories of associative learning. What happens during the interval that dramatically alters later fear memory?

Simulations of the Monfils-Schiller paradigm are shown in Figure 4.10. I simulated 3 conditions, differing only in the retrieval-extinction interval (REI): *No Ret* (REI=0), *Ret-short* (REI=3), *Ret-long* (REI=100).[6] As observed experimentally, all groups ceased responding by the end of extinction. Both Ret-long and No Ret showed spontaneous recovery after a long extinction-test delay. In contrast, Ret-short showed no spontaneous recovery at test. Examining the latent cause posteriors in the different conditions (Figure 4.10B-D), we see that the extinction trials were assigned to a new latent cause in the No Ret and Ret-long conditions, but to the training cause in the Ret-short condition.

During the retrieval-extinction interval, the CS-US association is reduced incrementally. This has the effect of making the CS-alone trials more likely under the training latent cause, since the prediction error decreases with each reduction of the

[6]Time is measured in arbitrary units here; see the Appendix for a description of how they roughly map onto real time.

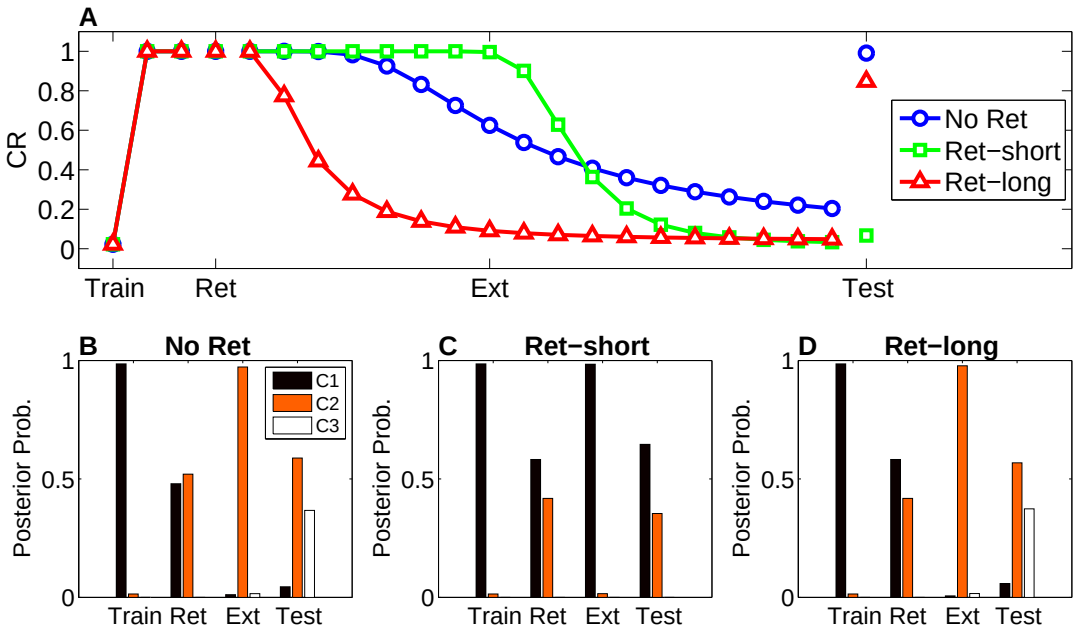

Figure 4.10: **Model predictions for the Monfils-Schiller paradigm**. Model predictions for the Monfils-Schiller paradigm. (A) Simulated conditioned response (CR) during training (3 CS-US pairs), Retrieval (Ret; 1 CS presentation 24 hours after training, followed by either a short or long interval), extinction (Ext; CS-alone presentations) and a test phase 24 hours later. Three conditions are shown: No-Ret (no retrieval trial), Ret-short (retrieval with a short post-retrieval interval), and Ret-long (retrieval with a long post-retrieval interval). (B-D) The posterior probability distribution over latent causes (denoted C1, C2 and C3) in each condition. Only the top 3 highest probability causes are shown here. The "ret" trial in the no-ret condition refers to the first trial of extinction.

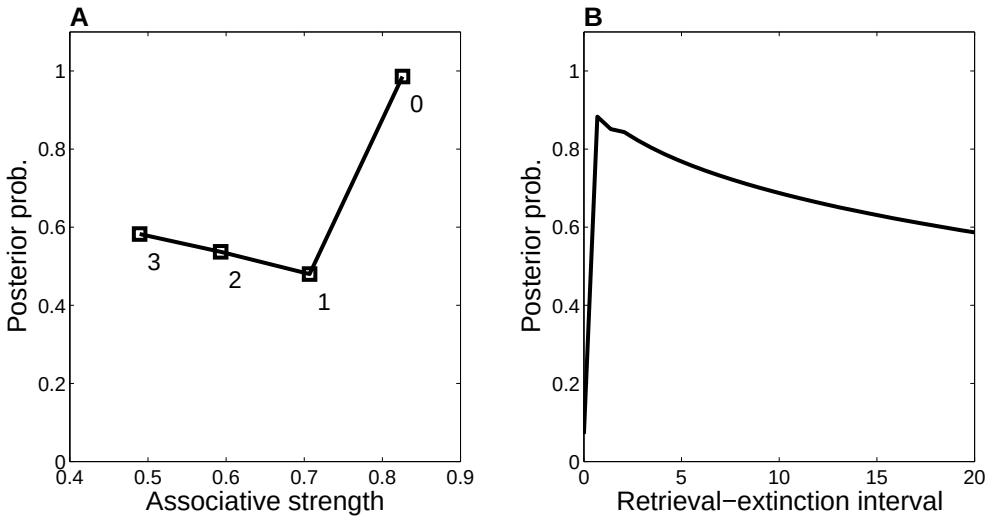

Figure 4.11: **Dynamics of associative and structure learning during the retrieval-test interval**. (A) The X-axis represents the associative weight corresponding to the training latent cause. The Y-axis represents the posterior probability that the training latent cause is active. Each numbered square indicates a particular iteration during the retrieval-test interval; the square numbered "0" indicates the timestep prior to retrieval. Initially, the prediction error causes the posterior to favor a new latent cause. Over the course of several iterations, incremental reductions in the associative weight pull the posterior probability higher by making the retrieval trial conditionally more likely under the training cause. (B) As the retrieval-extinction interval grows longer, the probability of assigning the first extinction trial to the training cause first peaks (due to incremental adjustment of the weights), then diminishes due to the time-sensitive prior (Eq. 4.3).

associative strength. Thus, in the Ret-short condition, the probability that the retrieval trial is assigned to the training latent cause increases over the course of the interval. Spontaneous recovery is attenuated due to the decrement of the training latent cause's CS-US association (Figure 4.11A). When the interval is too short (as in the No Ret condition), there is insufficient time (i.e., too few EM iterations) to reduce the CS-US association and tip the balance in favor of the training cause. When the retrieval-test interval is long (as in the Ret-long condition), the time-sensitive prior begins to exert a stronger effect, biasing the animal to assign the retrieval trial to a new latent cause. This nonmonotonic dependence on the retrieval-test interval is shown quantitatively in Figure 4.11B.

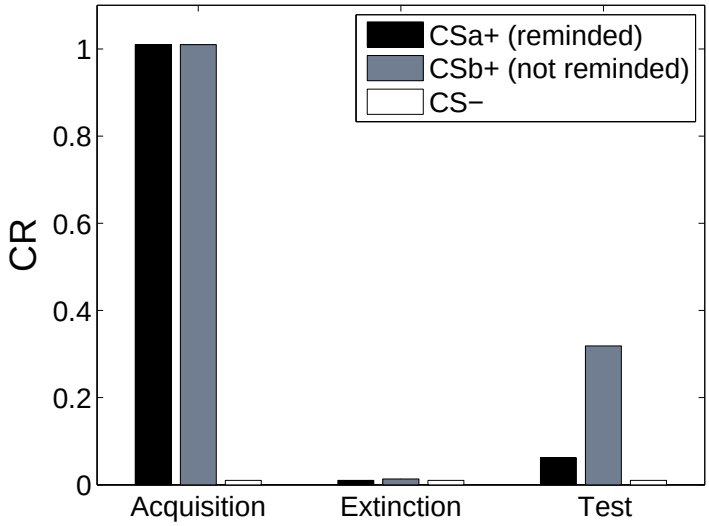

Figure 4.12: **Cue-specificity in the Monfils-Schiller paradigm**. Model simulations of the within-subjects design reported by Schiller et al. (2010), in which two CSs were trained separately, but only one was reexposed prior to extinction. Fear recovery is attenuated for the reexposed CS.

Figure 4.12 shows simulations of the cue-specificity experiment reported in Schiller et al. (2010). In a within-subjects design, two CSs were trained separately, but only one was reexposed prior to extinction. Consistent with the results of Doyère et al. (2007), Schiller et al. (2010) found that fear recovered for the CS that was not reexposed, but not for the reexposed CS. This finding fits with my theoretical interpretation that CS reexposure leads to memory modification for the US association specific to that CS and the reactivated latent cause.

The importance of iterative adjustment during the retrieval-test interval suggests that distracting or occupying animals during the interval should disrupt the Monfils-Schiller effect. For example, my theory predicts that giving rats a secondary task to perform during the interval will prevent unlearning of the CS-US association, leading to later recovery of fear. Alternatively, it might be possible to enhance the effect by leaving the animal in the conditioning chamber during the interval; the chamber would serve as a reminder cue, potentially preventing the animal from getting distracted.

4.5 Experiment: Performing Extinction Prior to Retrieval Attenuates Reconsolidation

My model predicts that subjects in the Monfils-Schiller paradigm retrieve and update the latent cause responsible for the training trials during the extinction session. This leads to the prediction that performing an extinction session prior to retrieval will render the paradigm ineffective in attenuating CS-US associations: The pre-retrieval extinction session will generate a new latent cause, which will subsequently be preferentially retrieved during the retrieval and post-retrieval extinction trials. In this case, extinction training will not modify the original memory trace (i.e., weaken the originally acquired CS-US association) because the acquisition cause was never assigned to the extinction or retrieval trials. At test, animals will be influenced by the original (unattenuated) association, and thus show fear, for the same reasons that animals recover fear after standard extinction.

To test this, I first fear-conditioned rats using 3 tone-shock pairings. On the next day, an "extinction-retrieval-extinction" (E-R-E) group of rats received a short extinction session (5 unreinforced tone CSs) while a "retrieval-extinction" (R-E) group did not undergo extinction prior to retrieval. Based on the predictions of my model, I hypothesized that the first group would infer a new latent cause for unreinforced trials, while the second would not, and that this would interact with the ability of a future retrieval+extinction session to modify the memory in the original training cause. To investigate this prediction, 24 hours later, rats were presented with an isolated retrieval cue (one non-reinforced tone CS), followed one hour later by an extinction session (18 unreinforced CSs). On the next day, all rats received 5 unsignaled footshocks, and 24 hours later they were tested for reinstatement of fear (Pavlov, 1927; Rescorla and Heth, 1975; Bouton and Bolles, 1979b). The experimental design is summarized in Figure 4.13A, and my model predictions are shown in Figure 4.13

(B-D).

4.5.1 Subjects

Eleven male Sprague-Dawley rats (250−300 g, Harlan Lab Animals Inc.) were used in this set of experiments. Procedures were conducted in compliance with the National Institutes of Health Guide for the Care and Use of Experimental Animals and were approved by the University of Texas at Austin Animal Care and Use Committee. Rats were housed in pairs in clear plastic cages and maintained on a 12-hour light/dark cycle with food and water provided ad libitum. Rats were handled for several minutes every day prior to the start of the experiment.

4.5.2 Apparatus and Stimuli

All behavioral procedures took place in standard conditioning chambers equipped with metal walls and stainless-steel rod floors connected to a shock generator and enclosed in acoustic isolation boxes (Coulbourn Instruments, Allentown, PA). Behavior was recorded using infrared digital cameras mounted on the top of each unit. The chambers were cleaned with Windex between sessions.

Stimulus delivery was controlled using Freeze Frame software (Coulbourn Instruments). A 20 second tone (5 kHz, 80 dB) played through a speaker in the walls of the box served as a conditional stimulus. The US was a 500 ms 0.7 mA foot-shock.

4.5.3 Behavioral Procedures

Fear conditioning. Rats were allowed to habituate to the chambers for 10 minutes before receiving three 20 second presentations of the tone [inter-trial intervals (ITIs) = 160s and 200s], each co-terminating with a foot-shock. After fear conditioning, all rats were returned to their home cage.

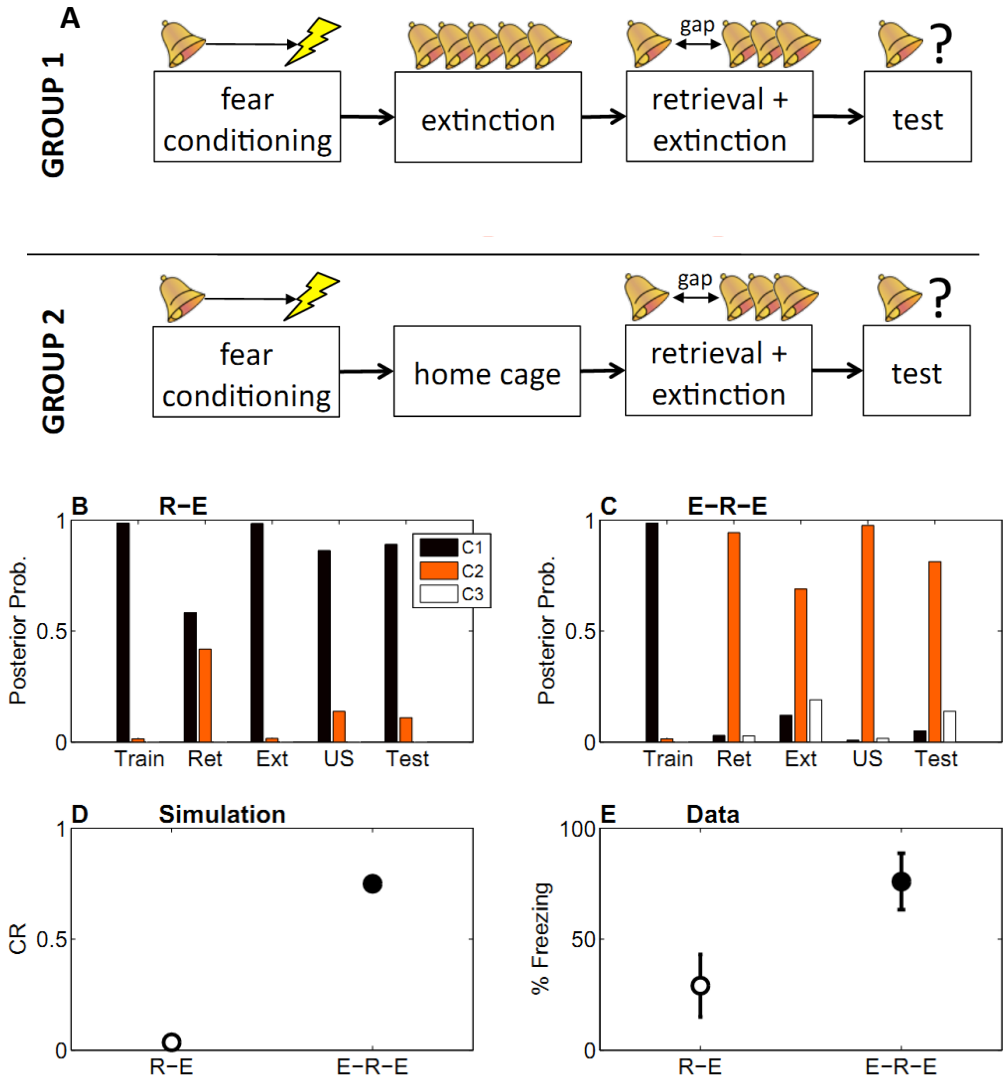

Figure 4.13: **Performing extinction prior to retrieval attenuates reconsolidation**. (A) Experimental design: Rats were fear-conditioned rats using 3 tone-shock pairings. On the next day, one group of rats (E-R-E) received an extinction session (5 non-reinforced tone CSs) while a second group (R-E) did not undergo extinction. Twenty-four hours later, rats received an isolated retrieval cue (one non-reinforced tone CS), followed one hour later by an extinction session (18 unreinforced CSs). On the next day, all rats received 5 unsignaled footshocks, and 24 hours later they were tested for reinstatement of fear. (B, C) Simulated latent cause posteriors for the two conditions. (D) Model predictions. (E) Experimental results.

Extinction. Twenty-four hours after fear conditioning, rats were divided into two groups (E-R-E and R-E). Rats in the E-R-E group received an 5 CS-alone presentation, while rats in the R-E group remained in the home cage. 24 hours later, both groups received an isolated CS presentation (retrieval trial), followed 1 hour later by 18 presentations of the tone in the absence of the foot-shock (ITI=160s). During the 1 hour interval, rats were returned to their home cage.

Reinstatement. Twenty-four hours after extinction, rats were returned to the chambers used for fear conditioning and extinction. The rats then received 2 unsignaled foot-shocks matched in intensity to the strength of the foot-shock administered during fear conditioning and extinction (0.7 mA) and were returned to their home cages upon completion. The next day, rats were returned to the experimental chamber and tested for reinstatement (4 tone presentations).

4.5.4 Scoring of Freezing Behavior

Freezing behavior was defined as the absence of any movement, excluding breathing and whisker twitching. The total number of seconds spent freezing throughout the tone presentation was expressed as a percentage of tone duration (20 seconds). Freezing was scored manually by an experimenter blind to group assignment.

4.5.5 Results

My experimental results (Figure 4.13E) are in accord with the model's predictions: rats that received an extinction session prior to the retrieval+extinction session showed greater reinstatement compared to the group that did not receive this initial extinction session. This constitutes a new boundary condition on reconsolidation: *the original fear memory is not updated when the retrieval trial is preceded by extinction.* My model anticipates this finding by assuming that the first extinction session leads to the inference of a new, "extinction" latent cause, rather than updating the acquisition

latent cause.

4.6 Discussion

I have shown how the major phenomena of memory reconsolidation can be accounted for by a rational analysis of Pavlovian conditioning. The key idea of this computational framework is a distinction between two learning processes: an associative learning process that adjusts the parameters of the animal's internal model, and a structure learning process that infers the latent causes underlying sensory inputs. I showed that the interplay between these two processes can reproduce parametric variations in the effects of reconsolidation treatments, consistent with experimentally observed boundary conditions.

One of the most intriguing reconsolidation findings in recent years was the discovery that a noninvasive behavioral treatment was effective at attenuating recovery (Monfils et al., 2009; Schiller et al., 2010). Monfils, Schiller and their colleagues demonstrated (in both rats and humans) that performing extinction training within a short interval following a retrieval cue (an unreinforced CS presentation) reduced recovery of fear. Ma et al. (2011) have recently demonstrated an analogous effect in an appetitive learning task. However, the effectiveness of this paradigm has been controversial, with several replication failures (Chan et al., 2010; Costanzi et al., 2011; Kindt and Soeter, 2011).

Recent work has leant biological plausibility to the claim that the Monfils-Schiller paradigm erases the CS-US association learned during training (Clem and Huganir, 2010). After fear conditioning, there is an upregulation of AMPA receptor trafficking to the post-synaptic membrane at thalamus-amygdala synapses, and memory is impaired if this trafficking is blocked (Rumpel et al., 2005), suggesting that changes in post-synaptic AMPA receptor density may be the neural substrate of associative

learning in fear conditioning. Clem and Huganir (2010) found that the Monfils-Schiller paradigm results in synaptic removal of calcium-permeable AMPA receptors; this finding is significant in that it indicates a reversal of the synaptic changes that occurred during training, supporting the view that the Monfils-Schiller paradigm results in unlearning of the CS-US association acquired during training.

My theoretical analysis is consistent with this view. I showed in simulations that during the retrieval-extinction interval, an associative learning process is engaged (and continues to be engaged during extinction training) that decrements the CS-US association, whereas standard extinction engages a structure learning process that assigns the extinction trials to a new latent cause, creating a new memory trace without modifying the original memory. This leads to the testable prediction that disrupting the neural substrates of associative learning, or potentiating the substrates of structure learning, during the retrieval-extinction interval should block memory updating in the Monfils-Schiller paradigm.

In a behavioral experiment, I examined another prediction of my computational framework: performing a retrieval trial after some extinction training has already taken place should be ineffective at preventing fear recovery. The reason is that the initial extinction trials will be assigned to a new, "extinction" latent cause, and post-retrieval extinction trials will then be assigned to this cause, in spite of the retrieval trial. As a consequence, the extinction cause will be retrieved rather than the acquisition cause, leading to fear recovery. My behavioral data confirm this prediction.

One challenge to developing a unified theory of reconsolidation is that some of the basic facts are still disputed. Some authors have found that contextual fear memories become labile after retrieval (Debiec et al., 2002), while others have not (Biedenkapp and Rudy, 2004), and yet others argue that the memory modification is transient (Frankland et al., 2006). A similar situation exists for instrumental memories. Some

studies have shown that instrumental memories undergo reconsolidation (Fuchs et al., 2009; Milton et al., 2008), while others have not (Hernandez and Kelley, 2004). There are many differences between these studies that could account for such discrepancies, including the type of amnestic agent, how the amnestic agent is administered (systemically or locally), the type of reinforcer, and the timing of stimuli. It would be hazardous to attempt a comprehensive theory of these phenomena before studies have been undertaken that isolate the critical experimental factors.

4.6.1 A Neural Circuit for Reconsolidation

Although I have so far not committed to any specific neural implementation of my model, I believe it fits comfortably into the computational functions of the circuit underlying Pavlovian conditioning. I propose a provisional mapping onto this circuit, centering on the amygdala and the "hippocampal-VTA loop" (Lisman and Grace, 2005) connecting the hippocampus and the ventral tegmental area in the midbrain. My basic proposal is inspired by two lines of research, one on the role of hippocampus in structure learning, and one on the role of the dopamine system and the amygdala in associative learning.

In previous work, I have suggested that the hippocampus is a key brain region involved in partitioning the world into latent causes (Gershman et al., 2010). This view resonates with earlier models emphasizing the role of the hippocampus in encoding sensory inputs into a statistically compressed latent representation (Fuhs and Touretzky, 2007; Gluck and Myers, 1993; Levy et al., 2005). Some of the evidence for this view comes from studies showing that context-specific memories depend on the integrity of the hippocampus (e.g., Honey and Good, 1993), indicating that animals without a hippocampus cannot "carve nature at its joints" (i.e., partition observations into latent causes; see Gershman and Niv, 2010).

Within the current model, I propose that the dentate gyrus (DG) activates la-

tent representations of the sensory inputs in area CA3. Each of these representations corresponds to a latent cause, and their level of activation is proportional to their prior probability (Eq. 4.3). Mechanistically, these representations may be encoded in attractors by the dense recurrent collaterals that are distinctive of CA3 (McNaughton and Morris, 1987). An important aspect of my model is that the repertoire of latent causes can expand adaptively. One potential mechanism for creating new attractors is neurogenesis of granule cells in the DG (Becker, 2005). This account predicts that the role of neurogenesis in creating new attractors should be time-sensitive in a manner comparable to the latent cause prior (i.e., implement the contiguity principle). Consistent with this hypothesis, Aimone et al. (2006) have suggested that immature granule cells, by virtue of their low activation thresholds, high resting potentials and constant turnover, cause inputs nearby in time to map onto the same CA3 representation.

There is widespread agreement that CS-US associations in auditory fear conditioning are encoded by synapses between the thalamus and the basolateral amygdala (BLA; McNally et al., 2011). Accordingly, I suggest that the amygdala transmits a US prediction that is then compared to sensory afferents from the periacqueductal gray region of the midbrain. The resultant prediction error is computed in the ventral tegmental area (VTA) and transmitted by dopaminergic projections to both the amygdala and CA1.

The role of dopamine in associative learning is well established (see Glimcher, 2011, for a recent review), and has been specifically implicated in Pavlovian fear conditioning (Pezze and Feldon, 2004), although little is known about the phasic firing properties of dopamine neurons during fear conditioning. Dopamine gates synaptic plasticity in the BLA (Bissière et al., 2003), consistent with its hypothesized role in driving the learning of CS-US associations. I hypothesize that dopaminergic inputs to CA1 reflect the influence of reward prediction errors on the posterior distribution over

latent causes. The output of CA1 feeds back into the VTA by way of the subiculum (Lisman and Grace, 2005), potentially providing a mechanism by which the posterior can modulate the prediction errors, as predicted by my model.

4.6.2 Comparison to Other Models

Several other theoretical frameworks have been proposed to account for various aspects of reconsolidation. In this section, I briefly describe two and compare them to my own.

Osan et al. (2011)

Osan et al. (2011) have proposed an autoassociative neural network model of reconsolidation that explains many of the reported boundary conditions in terms of attractor dynamics (see also Amaral et al., 2008, for a related model). In this model, training and extinction memories correspond to attractors in the network, formed through Hebbian learning. Given a configuration of sensory inputs, the state of the network evolves towards one of these attractors. In addition, a "mismatch-induced degradation" process adjusts the associative weights that are responsible for the mismatch between the retrieved attractor and the current input pattern; this mismatch is assumed to accumulate over the course of the input presentation. The degradation process, in the case of extinction, implements a form of unlearning. The relative balance of Hebbian learning and mismatch-induced degradation determines the outcome of extinction training. Administration of PSIs (e.g., anisomycin) is modeled by a scalar factor that downweights the influence of Hebbian plasticity in the weight updates.

Osan et al. (2011) showed that their network model could account for a number of the boundary conditions on reconsolidation described above. For example, they simulated the effect of CS reexposure duration prior to PSI administration (Eisenberg

et al., 2003; Suzuki et al., 2004): On very short reexposure trials, the shock memory is preferentially retrieved because it has already been encoded in an attractor as a consequence of training (i.e., the shock memory is the dominant trace). The accumulated mismatch is small, and hence mismatch-induced degradation has little effect on the shock memory. Since the mismatch is close to zero and the effect of PSIs is to turn off Hebbian learning, the net effect of PSI administration following reexposure is no change in the memory. On long reexposure trials, the accumulated mismatch becomes large enough to favor the formation of a new attractor corresponding to the extinction memory (i.e., the no-shock memory is the dominant trace). In this case, PSI administration will have no effect on the shock memory, because Hebbian learning is operating on a different attractor.

Post-reexposure PSI administration has a tangible effect on the shock memory for intermediate durations (i.e., what I modeled as "short" duration in my simulations of the PSI experiments). In this case, mismatch is large enough to induce degradation, but not large enough to induce the formation of a new attractor. The PSI prevents Hebbian learning from compensating for this degradation by modifying the associative weights. Regardless of whether modification happens through Hebbian learning or mismatch-induced degradation, the important point here is that the shock memory is modified, rather than a new attractor being formed.

In addition to the parametric effect of reexposure duration on reconsolidation, Osan et al. (2011) also simulated the effects of memory strength (more highly trained memories are resistant to labilization by PSI administration), the effects of NMDA receptor agonists (which have the opposite effects of PSIs), and the effects of blocking mismatch-induced degradation (the amnestic effect of PSI administration is attenuated). However, the model of Osan et al. (2011) is fundamentally limited by the fact that it lacks an explicit representation of time. This prevents it from accounting for the results of the Monfils-Schiller paradigm: all the retrieval-extinction inter-

vals should lead to the same behavior (contrary to the empirical data). The lack of temporal representation also prevents it from modeling the effects of memory age on reconsolidation, since there is no mechanism for taking into account the interval between training and reexposure. In contrast, the latent cause model explicitly represents temporal distance between observations, making it sensitive to changes in timing.[7]

Another problem with the model of Osan et al. (2011) is that in order to explain spontaneous recovery, it was necessary to introduce an ad hoc function that governs pattern drift during reexposure. This function—by construction—produces spontaneous recovery, but it is not obvious why pattern drift should follow such a function. No psychological or neurobiological justification is provided.

One appealing feature of the Osan et al. (2011) model is its neurobiological plausibility. We know that attractor networks exist in the brain (e.g., in area CA3 of the hippocampus), and (in certain circumstances) support the kinds of learning described above. The model provides a simplified but plausible mapping from computational variables to biological substrates. As I discussed in the previous section, one way to think about latent causes at a neural level is in terms of attractors (e.g., in area CA3). Thus, although the formal details of Osan et al. (2011) differ from my own, there may be neural implementations of the latent cause model that bring it closer to the formalism of the attractor network. However, in its current form the latent cause model is not specified at the same biologically detailed level as the model of Osan et al. (2011); it makes no distinction between Hebbian plasticity and mismatch-induced degradation, and consequently has nothing to say about pharmacological manipulations that selectively effect one or the other process, for example the disruption of mismatch-induced degradation by inhibitors of the ubiquitin-proteasome cascade (Lee et al., 2008).

[7]Conceivably, one could incorporate a time-sensitive mechanism by using a "temporal context" vector of the sort described in Chapter 8 as part of the input patterns.

Sederberg et al. (2011)

The model of Sederberg et al. (2011) was developed to explain a set of experimental findings reported by Hupbach et al. (2007, 2009); see Chapter 8 for more details. Using a list-learning paradigm with humans, Hupbach et al. (2007) showed that reminding participants of one list (A) shortly before giving them a second list (B) to study produced an asymmetric pattern of intrusions at test: participants intruded a large number of items from list B when asked to recall list A, but not vice versa. When no reminder was given, participants showed an overall low level of intrusions across list A and list B recall. Sederberg et al. (2011) proposed a variant of the Temporal Context Model (TCM; Howard and Kahana, 2002) to account for these findings. The basic idea underlying TCM is that studied items are bound to a gradually drifting representation of temporal context—a recency-weighted average of previously experienced items. TCM explains the asymmetric pattern of intrusions in terms of the structure of item-context associations: the reminder treatment causes list B items to be associated to both the list A and list B contexts, whereas list A items are associated only with the list A context.

While TCM differs in many ways from my model, it shares the property that sensory inputs experienced in similar temporal contexts should be effectively clustered together. However, an important property that sets TCM apart from my model is that its explanation of reconsolidation is fundamentally retrieval-focused: list B items do not overwrite list A items following a reminder, but instead bias later retrieval. This position is representative of a large class of memory models that attribute the causes of forgetting to retrieval interference rather than memory decay or erasure (see Norman et al., 2006, for a review). At present, the superiority of a retrieval- or storage-focused interpretation of reconsolidation is hotly contested (Nader and Hardt, 2009; Riccio et al., 2006); I will not attempt to resolve this debate here, except to say that I have staked out one possible theoretical position that involves both storage

and retrieval processes.

4.6.3 Conclusion

The phenomenon of reconsolidation presents a particularly staunch challenge to contemporary theories of learning and memory. In this chapter I have attempted to comprehensively address this phenomenon from a rational Bayesian perspective. The mechanistic implementation of my rational analysis yields a new set of computational ideas with which to understand learning in Pavlovian conditioning and beyond. In particular, I have suggested that the interplay between associative and structure learning has momentous consequences for the fate of memory traces. By taking a computational approach, I can begin to harness this interplay and direct it towards modifying maladaptive memories such as trauma and addiction.

4.7 Appendix: computational model details

In this section, I provide the mathematical and implementational details of my model.

4.7.1 The expectation-maximization algorithm

The EM algorithm, first introduced by Dempster et al. (1977), is a method for performing maximum-likelihood parameter estimation in latent variable models. In my model, the latent variables correspond to the vector of latent cause assignments, $\mathbf{z}_{1:t}$, the parameters correspond to the associative weights, $\mathbf{W}$, and the data correspond to the history of cues and rewards, $\mathcal{D}_{1:t} = \{\mathbf{X}_{1:t}, \mathbf{r}_{1:t}\}$, where $\mathbf{X}_{1:t} = \{\mathbf{x}_1, \ldots, \mathbf{x}_t\}$ and $\mathbf{r}_{1:t} = [r_1, \ldots, r_t\}$. Let $Q(\mathbf{z}_{1:t}]$ be a distribution over $\mathbf{z}_{1:t}$. The EM algorithm can be

understood as performing coordinate ascent on the functional

$$\begin{aligned}\mathcal{F}(\mathbf{W}, Q) &= \sum_{\mathbf{z}_{1:t}} Q(\mathbf{z}_{1:t}|\mathcal{D}_{1:t}) \log P(\mathbf{z}_{1:t}, \mathcal{D}_{1:t}|\mathbf{W}) \\ &= \sum_{\mathbf{z}_{1:t}} Q(\mathbf{z}_{1:t}|\mathcal{D}_{1:t}) \log \left[P(\mathcal{D}_{1:t}|\mathbf{z}_{1:t}, \mathbf{W}) P(\mathbf{z}_{1:t})\right]. \end{aligned} \tag{4.10}$$

By Jensen's inequality, this functional is a lower bound on the log marginal likelihood of the data, $\log P(\mathcal{D}_{1:t}|\mathbf{W}) = \log \sum_{\mathbf{z}_{1:t}} P(\mathcal{D}_{1:t}, \mathbf{z}_{1:t}|\mathbf{W})$, which means that maximizing $\mathcal{F}$ corresponds to optimizing the internal model to best predict the observed data (Neal and Hinton, 1998).

The EM algorithm alternates between maximizing $\mathcal{F}(\mathbf{W}, Q)$ with respect to $\mathbf{W}$ and Q. Letting n indicate the iteration,

$$\begin{aligned}\textbf{E-step} &: Q^{n+1} \leftarrow \arg\max_{Q} \mathcal{F}(\mathbf{W}^n, Q) \\ \textbf{M-step} &: \mathbf{W}^{n+1} \leftarrow \arg\max_{\mathbf{W}} \mathcal{F}(\mathbf{W}, Q^{n+1})\end{aligned}$$

Alternating the E and M steps repeatedly, $\mathcal{F}(\mathbf{W}, Q)$ is guaranteed to converge to a local maximum (Neal and Hinton, 1998). It can also be shown that $\mathcal{F}(\mathbf{W}, Q)$ is maximized with respect to $Q(\mathbf{z}_{1:t})$ when $Q = P(\mathbf{z}_{1:t}|\mathcal{D}_{1:t}, \mathbf{W})$. Thus, the optimal E-step is exact Bayesian inference over the latent variables $\mathbf{z}_{1:t}$.

There are two challenges facing a biologically and psychologically plausible implementation of this algorithm. First, the E-step is intractable, since it requires summing over an exponentially large number of possible latent cause assignments. Second, both steps involve computations operating on the entire history of observations, whereas a more plausible algorithm is one that operates online, one observation at a time (Anderson, 1990). Below I summarize an approximate, online form of the algorithm. To reduce notational clutter, I drop the n superscript (indicating EM iteration), and implicitly condition on $\mathbf{W}$.

4.7.2 The E-step: structure learning

The E-step corresponds to calculating the posterior using Bayes' rule (Eq. 4.5). The number of terms in the summation over $\mathbf{z}_{1:t-1}$ grows exponentially over time; consequently, calculating the posterior exactly is intractable. Following Anderson (1991), I use a "local" *maximum a posteriori* (MAP) approximation (see Sanborn et al., 2010, for more discussion):

$$q_{tk} \approx \frac{P(\mathcal{D}_t|z_t = k, \hat{\mathbf{z}}_{1:t-1}, \mathcal{D}_{1:t-1})P(z_t = k|\hat{\mathbf{z}}_{1:t-1})}{\sum_j P(\mathcal{D}_t|z_t = j, \hat{\mathbf{z}}_{1:t-1}, \mathcal{D}_{1:t-1})P(z_t = j|\hat{\mathbf{z}}_{1:t-1})}, \tag{4.11}$$

where $\hat{\mathbf{z}}_{1:t-1}$ is defined recursively according to:

$$\hat{z}_t = \arg\max_k P(\mathcal{D}_t|z_t = k, \hat{\mathbf{z}}_{1:t-1}, \mathcal{D}_{1:t-1})P(z_t = k|\hat{\mathbf{z}}_{1:t-1}). \tag{4.12}$$

In other words, the local MAP approximation is obtained by replacing the summation over partitions with the sequence of conditionally optimal cluster assignments. Although this is not guaranteed to arrive at the globally optimal partition (i.e., the partition maximizing the posterior over all timepoints), in my simulations it tends to produce very similar solutions to more elaborate approximations like particle filtering (Gershman and Niv, 2010; Sanborn et al., 2010).[8]

The first term in Eq. 4.12 (the likelihood) is derived using standard results in Bayesian statistics (Bishop, 2006):

$$P(\mathcal{D}_t|z_t = k, \hat{\mathbf{z}}_{1:t-1}, \mathcal{D}_{1:t-1}) = \mathcal{N}(r_t; \hat{r}_{tk}, \sigma_r^2) \prod_{d=1}^{D} \mathcal{N}(x_{td}; \hat{x}_{tkd}, \nu_{tk}^2), \tag{4.13}$$

[8]The local MAP approximation has also been investigated in the statistical literature. Wang and Dunson (2011) found that it compares favorably to fully Bayesian inference, while being substantially faster.

where

$$\hat{r}_{tk} = \sum_{d=1}^{D} x_{td} w_{kd} \tag{4.14}$$

$$\hat{x}_{tkd} = \frac{N_{tk}\bar{x}_{tkd}}{N_{tk} + \sigma_x^2} \tag{4.15}$$

$$\nu_{tk}^2 = \frac{\sigma_x^2}{N_{tk} + \sigma_x^2} + \sigma_x^2. \tag{4.16}$$

Here N_{tk} denotes the number of times $z_\tau = k$ for $\tau < t$ and $\bar{x}_{tkd}$ denotes the average cue values for observations assigned to cause k for $\tau < t$. The second term in Eq. 4.12 (the prior) is given by the time-sensitive Chinese restaurant process (Eq. 4.3).

4.7.3 The M-step: associative learning

The M-step is derived by differentiating $\mathcal{F}$ with respect to $\mathbf{W}$ and then taking a gradient step to increase the lower bound. This corresponds to a form of stochastic gradient ascent, and is in fact remarkably similar to the Rescorla-Wagner learning rule (see below). Its main departure lies in the way it allows the weights to be modulated by a potentially infinite set of latent causes. Because these latent causes are unknown, the animal represents an approximate distribution over causes, $\mathbf{q}$ (computed in the E-step). The components of the gradient are given by:

$$[\nabla \mathcal{F}]_{kd} = \sigma_r^{-2} x_{td} \delta_{tk}, \tag{4.17}$$

where δ_{tk} is given by Eq. 4.4. To make the similarity to the Rescorla-Wagner model clearer, I absorb the σ_r^{-2} factor into the learning rate, η.

4.7.4 Simulation parameters

With two exceptions, I used the following parameter values in all the simulations: $\alpha = 0.1, \eta = 0.3, \sigma_r^2 = 0.4, \sigma_x^2 = 1, \theta = 0.02, \lambda = 0.01$. For modeling the retrieval-extinction data, I treated θ and λ as free parameters, which I fit using least-squares. For simulations of the human data in Figure 4.12, I used $\theta = 0.0016$ and $\lambda = 0.00008$. Note that θ and λ change only the scaling of the predictions, not their direction; all ordinal relationships are preserved.

The CS was modeled as a unit impulse: $x_{td} = 1$ when the CS is present and 0 otherwise (similarly for the US). Intervals of 24 hours were modeled as 20 time units; intervals of one month were modeled as 200 time units. While the choice of time unit was somewhat arbitrary, my results do not depend strongly on these particular values.

4.7.5 Relationship to the Rescorla-Wagner model

In this section I demonstrate a formal correspondence between the classic Rescorla-Wagner model and my model. In the Rescorla-Wagner model, the outcome prediction $\hat{r}_t$ is, as in my model, parameterized by a linear combinations of the cues $\mathbf{x}_t$ and is updated according to the prediction error:

$$\hat{r}_t = \sum_{d=1}^{D} w_d x_{td} \tag{4.18}$$

$$\delta_t = r_t - \hat{r}_t \tag{4.19}$$

$$\mathbf{w} \leftarrow \mathbf{w} + \eta \mathbf{x}_t \delta_t. \tag{4.20}$$

The key difference is that in my model, I allow there to be separate weight vectors for each latent cause. When $\alpha = 0$, the distribution over latent causes reduces to a delta function at a single cause (since the probability of inferring new latent causes is

always 0), and hence there is only a single weight vector. In this case, the two models coincide.

Chapter 5

Gradual extinction in Pavlovian fear conditioning

Once a fear memory trace is laid down in the brain, can it be erased? As described in Chapter 1, when animals are conditioned to associate a cue with an aversive stimulus, repeatedly presenting the cue alone (extinction training) reduces their fear of the cue. Unfortunately, this reduction is temporary, and fear generally returns with the passage of time, a phenomenon known as spontaneous recovery (Pavlov, 1927; Rescorla, 2004). Fear also generally returns following an isolated occurrence of the aversive stimulus, a phenomenon known as reinstatement (Pavlov, 1927; Rescorla and Heth, 1975; Bouton and Bolles, 1979b). Rather than modifying the fear memory, it is believed that extinction training creates a new memory that only transiently inhibits the original association (Bouton, 1993).

The onset of extinction training produces a large prediction error—a discrepancy between the predicted outcome (e.g., shock) and the experienced outcome (no shock). Traditional models of associative learning propose that such prediction errors serve as a learning signal, driving the modification of predictions (e.g., Rescorla and Wagner, 1972). According to these accounts, the absence of shocks during the extinction

procedure should reduce the strength of the original fear memory. However, recent models (such as the ones described in Chapters 3 and 4) propose that persistently large prediction errors might also serve as a segmentation signal, indicating to the animal a novel situation that demands new associations (Redish et al., 2007; Gershman et al., 2010). This can explain why the traditional extinction procedure leads to formation of a new, competing, "no-fear" memory, all the while allowing the original fear memory to persist unmodified.

The idea that large prediction errors are a signal for segmentation suggests that one could modify the original fear memory if prediction errors were small or infrequent enough to not induce formation of a new memory, but large enough to drive some learning. To test this prediction, I designed a "gradual extinction" paradigm in which the aversive event (a foot shock) was gradually and progressively eliminated. The idea was to change the association of the cue from a shock to no shock gradually enough so as to avoid persistent, large prediction errors. If we could prevent the creation of a new memory trace, all learning would affect the old fear memory, which would gradually be weakened and erased.

5.1 Methods

Subjects

Seventy-nine male Sprague-Dawley rats ($250 - 300$ g; Harlan Lab Animals Inc.) were used in this set of experiments. Forty-seven rats were used in Experiment 1 (16 in the Standard and Gradual groups, 15 in the Gradual Reverse group), and 32 were used in Experiment 2 (12 in each of the Reverse and Gradual groups, 8 in the Standard group). Procedures were conducted in compliance with the National Institutes of Health Guide for the Care and Use of Experimental Animals and were approved by the University of Texas at Austin Animal Care and Use Committee. Rats were housed

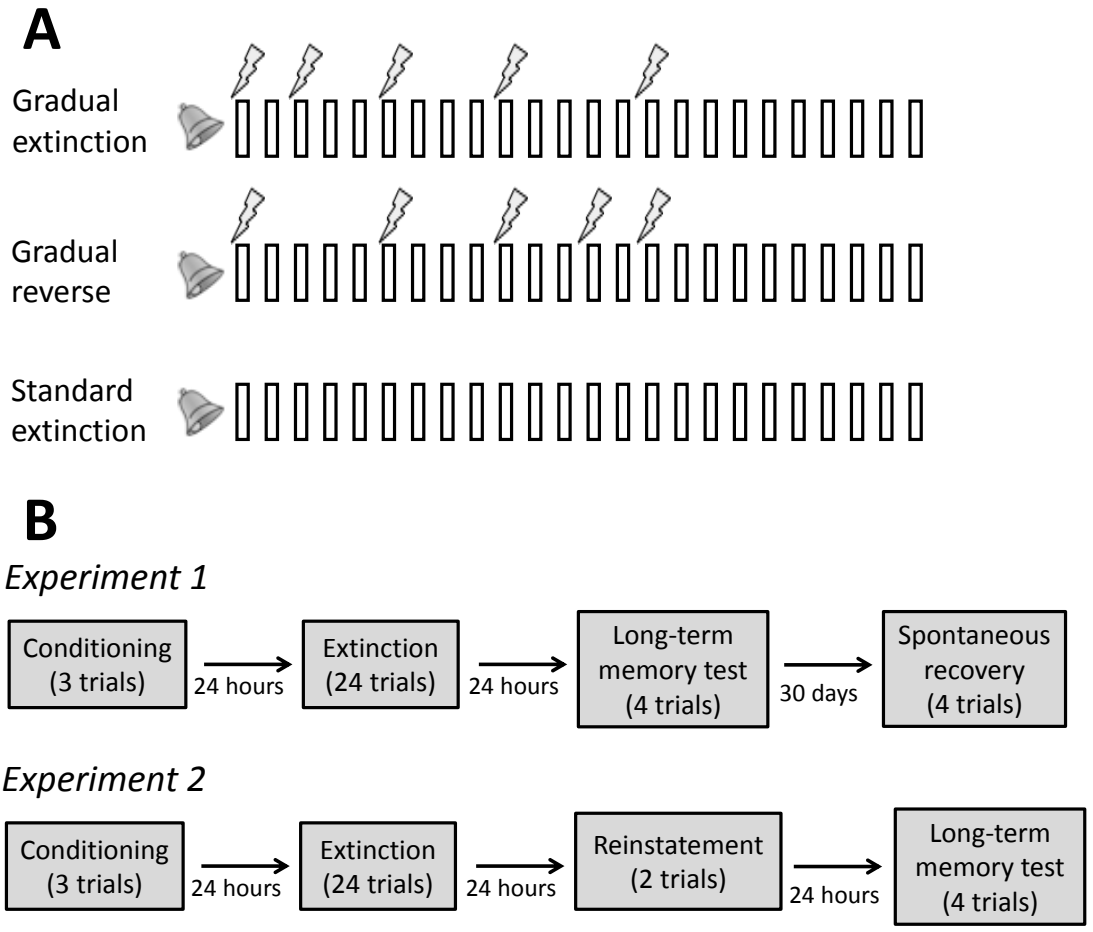

Figure 5.1: (A) Schematic of the extinction phase in each extinction condition. Bars represent 20 second tone presentations; lightning bolts represent 500 ms 0.7 mA foot shocks. (B) Design of Experiments 1 and 2.

in pairs in clear plastic cages and maintained on a 12-hour light/dark cycle with food and water provided ad libitum. Rats were handled for several minutes every day prior to the start of the experiment.

Apparatus and stimuli

All behavioral procedures took place in standard conditioning chambers equipped with metal walls and stainless-steel rod floors connected to a shock generator and enclosed in acoustic isolation boxes (Coulbourn Instruments, Allentown, PA). Behavior was recorded using infrared digital cameras mounted on the top of each unit. The chambers were cleaned with Windex between sessions.

Stimulus delivery was controlled using Freeze Frame software (Coulbourn Instruments). A 20 second tone (5 kHz, 80 dB) played through a speaker in the walls of the box served as a conditional stimulus. The unconditional stimulus was a 500 ms 0.7 mA foot-shock.

Behavioral procedures

Fear conditioning. Rats were allowed to habituate to the chambers for 10 minutes before receiving three 20 second presentations of the tone [inter-trial intervals (ITI) = 160s and 200s], each co-terminating with a foot-shock. After fear conditioning, all rats were returned to their home cage.

Extinction. Twenty-four hours after fear conditioning, all rats were divided into three groups (Standard, Gradual, and Reverse). Rats in the Standard group received 24 presentations of the tone in the absence of the foot-shock. Rats in the Gradual group also received 24 tone presentations; trials 1, 3, 6, 10, and 15 were paired with a foot-shock, resulting in a gradual decrease in the frequency of the shock. Rats in the Gradual Reverse group received 24 tone presentations with trials 1, 6, 10, 13, 15 paired with a foot-shock, resulting in a gradual increase in the frequency of the shock.

For all groups the last 9 trials included only tones with no shocks. After extinction, rats were returned to their home cage. All ITIs were 160s.

Long-Term Memory Test (Experiment 1). Twenty-four hours after extinction, rats were tested for long-term memory of the extinction phase by recording freezing during four presentations of the tone.

Spontaneous Recovery Test (Experiment 1). Thirty days after extinction, rats were returned to the chambers for a test of spontaneous recovery of fear by recording freezing during four presentations of the tone.

Reinstatement Test (Experiment 2). Twenty-four hours after extinction, rats were returned to the chambers used for fear conditioning and extinction. The rats then received 2 unsignaled foot-shocks and were returned to their home cages upon completion. The next day, rats were returned to the experimental chamber and tested for reinstatement of fear by recording freezing during four presentations of the tone.

Scoring of Freezing Behavior. Freezing behavior was defined as the absence of any movement, excluding breathing and whisker twitching. The total number of seconds spent freezing throughout each tone presentation was expressed as a percentage of tone duration (20 seconds). Freezing was scored manually by an experimenter blind to group assignment.

5.2 Results

After a conditioning phase in which the foot shock was paired with a tone three times, rats received 24 tone presentations in an extinction phase; however, five of these trials (trials 1, 3, 6, 10 and 15) co-terminated with a shock, such that the frequency of shocks decreased gradually (Figure 5.1A). I compared this gradual extinction schedule to both a standard extinction control condition (in which no shocks were presented in the extinction phase), and to a gradual reverse control condition in which five shocks

were presented in the extinction phase, but at a gradually increasing rather than decreasing frequency. To ensure that all groups extinguished to the same level, the last 9 tones were always presented without shock (Figure 5.1A).

I predicted that in both control groups large and persistent prediction errors at the beginning of extinction would induce formation of a new memory and prevent the modification of the old memory, thus fear would not be erased and would ultimately return. In the gradual extinction condition, in contrast, I predicted a permanent reduction of fear due to the non-reinforced trials modifying the original fear memory. To assess the persistence of fear memory, in Experiment 1 I tested for spontaneous recovery of fear and in Experiment 2 I tested for reinstatement (Figure 5.1B).

In Experiment 1 all three groups (n = 16 for the Gradual and Standard groups, $n = 15$ for the Gradual Reverse group) showed equivalent levels of freezing on the last four trials of extinction (one-way ANOVA, $P = 0.502$; Figure 5.2A), indicating similar degrees of fear of the tone. I then tested for the return of fear using a spontaneous recovery test one month following extinction. To measure spontaneous recovery, I calculated the difference between freezing on the first 4 trials of the spontaneous recovery test and the last 4 trials of extinction. There was a significant effect of group on freezing on this difference score [one-way ANOVA, $F(1, 44) = 4.26, P < 0.05$; Figure 5.2B]. A planned contrast $[Gradual - (Standard + GradualReverse)]$ within the ANOVA showed that the difference score for the Gradual group was significantly less than for the Standard and Gradual Reverse groups $[F(1, 44) = 8.32, P < 0.01]$. The raw measures of freezing in the spontaneous recovery test showed the same pattern: there was a significant effect of group [one-way ANOVA, $F(1, 44) = 3.26, P < 0.05$] and planned comparisons showed that rats in the Gradual group froze significantly less than in the Standard and Gradual Reverse group $[F(1, 44) = 5.51, P < 0.05]$.

In Experiment 2 I employed a reinstatement design: 24 hours after extinction, rats ($n = 12$ for the Gradual and Gradual Reverse groups, $n = 8$ for the Standard group)

were exposed to 2 unsignaled shocks, and then tested 24 hours later for freezing to the tone in a reinstatement test (see Methods). Figure 5.2C shows freezing during extinction and the reinstatement test. As in Experiment 1, a one-way ANOVA found no significant difference in freezing between the groups on the last four trials of extinction ($P = 0.07$). To measure reinstatement, I calculated the difference between freezing on the 4 trials of the reinstatement test and the last 4 trials of extinction. There was a significant effect of group on freezing on this difference score [one-way ANOVA, $F(1, 29) = 6.70, P < 0.005$; Figure 5.2D]. A planned comparison showed that the difference score for the Gradual group was significantly less than for the Standard and Gradual Reverse groups [$F(1, 29) = 13.13, P < 0.005$]. Here too the raw measure of freezing showed a similar pattern: there was a significant effect of group on freezing in the reinstatement test [one-way ANOVA, $F(1, 29) = 4.04, P < 0.05$] and planned comparisons showed that rats in the Gradual group froze significantly less than in the Standard and Gradual Reverse group [$F(1, 29) = 7.94, P < 0.01$].

To ensure that my results were not an artifact of pre-tone freezing, I confirmed that freezing measured during the 20 seconds prior to the first tone presentation in the extinction session was minimal and was not significantly differently between groups in any of the experiments (one-way ANOVA; $P = 0.152$ and $P = 0.866$, for Experiments 1 and 2, respectively). In Experiment 1, the pre-tone freezing measured before the SR test was also minimal, and was not significantly different between groups (one-way ANOVA, $P = 0.126$).

5.3 Discussion

In two fear conditioning experiments with rats, I found that gradually reducing the tone-shock contingency during extinction was effective in preventing the subsequent return of fear. This is in contrast to regular extinction protocols that transition

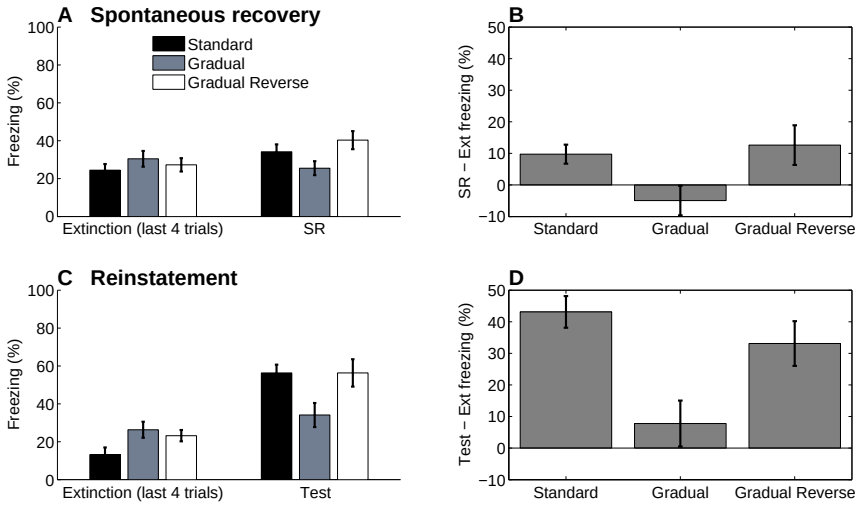

Figure 5.2: **Results of Experiments 1 and 2**. The left panels show freezing on the last 4 trials of extinction and at test, the right panels show the difference score (percent freezing) between the test phase and the end of extinction. Error bars represent standard error of the mean. (A,B) Results of Experiment 1, in which animals were tested for spontaneous recovery of fear 1 month after extinction. Freezing 30 days after extinction (SR test) was greater than freezing on the last four trials of extinction (Ext) in the Standard and Gradual Reverse groups compared to the Gradual group. (C,D) Results of Experiment 2, in which animals were exposed to 2 unsignaled shocks 24 hours after extinction, followed by a reinstatement test 24 hours later. On the reinstatement test the Standard and Gradual Reverse groups froze significantly more than the Gradual group.

abruptly from reinforced to non-reinforced presentations of the cue, and in general are ineffective at permanently extinguishing the conditioned response (e.g., fear). Importantly, my results cannot be simply attributed to partial reinforcement during extinction: a Gradual Reverse control condition, in which the tone and shock were paired the same number of times as in the Gradual condition but with increasing frequency, led to the return of fear. Thus, the outcome of fear extinction depends in subtle ways on the precise schedule of reinforcement. This pattern of dependence was predicted by my new theory of associative learning, which served as the impetus for these experiments.

My results fit well with an emerging set of theoretical ideas (see Chapters 3 and 4) that generalize single-association models such as in Rescorla and Wagner (1972). In a single-association model (Figure 5.3A), one association is learned for each cue-reinforcer pair. Such models, the mainstay of traditional associative learning theory, typically have trouble dealing with fear recovery phenomena (though see Schmajuk et al., 1996), due to the fact that during extinction they unlearn the association acquired during conditioning. In contrast, multiple-association models allow a cue to activate different associations at different times: if one association is activated in acquisition and another in extinction, the acquisition association remains intact and can result in resurgence of fear as in spontaneous recovery and reinstatement.

In Chapter 4, I suggested that multiple associations arise from animals inferences about the latent causes that give rise to their sensory data (see also Redish et al., 2007). According to this theory, when conditions change considerably (such as when transitioning from acquisition to extinction), the animal infers that a new latent cause is responsible for the observed data, and creates a new association. In this way, each latent cause is manifest in a separate associative weight, and inference about which latent causes are active modulates the effect of each associative weight on the prediction of the reinforcer (and thus the conditioned response; Figure 5.3B). That is,

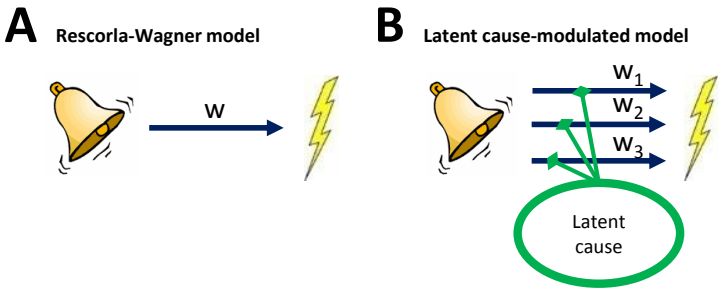

Figure 5.3: (A) The associative structure underlying the Rescorla-Wagner model. A learnable associative strength between a conditional stimulus (tone) and an unconditional stimulus (shock) is encoded by a scalar weight, w. (B) The associative structure underlying the latent cause-modulated model. As in the Rescorla-Wagner model, associative strength is encoded by a scalar weight, but in this case there is a collection of such weights, each paired with a different latent cause. The shock prediction is a linear combination of weights, modulated by the posterior probability that the corresponding latent cause is active.

if the animal infers that the latent cause active at test is the same as that which was active in acquisition, the related associative weight will generate a strong prediction of a forthcoming shock, and thus a fear response.

Importantly, in my theory the animals belief about whether a particular latent cause is active is determined by the similarity between the current situation and those that occurred when the latent cause was previously active. This explains why abrupt extinction, in which conditions change dramatically, brings about inference of a new latent cause and learning of a new associative weight (rather than modification of the original acquisition association). By titrating the similarity between extinction and acquisition, and only gradually moving away from the acquisition scenario, I endeavored to prevent the inference of a new latent cause, and instead continue modifying (and gradually erasing) the acquisition association. My results suggest that the manipulation was successful in doing just that. My theory thus provides a computational framework for understanding the interplay between learning and memory: A memory trace summarizes the statistical properties of a set of experiences (e.g., cue-reinforcer pairings) that have been assigned to a single latent cause, and learn-

ing occurs when a new experience is assigned to the same latent cause. If the new experience is slightly different from the old ones, its assignment to the same latent cause results in modification of the memory trace associated with the latent cause. In contrast, a new memory trace is formed when new experience fails to match the statistics of any existing memory traces.

Following the model presented in Chapter 4, I propose that persistently large, frequent prediction errors lead to new memory formation. This is because prediction errors arise when current experience is highly dissimilar to what is expected based on previous experience. Neurally, this process might rely on an interaction between dopaminergic prediction error signals originating in the midbrain (Schultz et al., 1997) and hippocampal pattern separation processes in the dentate gyrus (Lisman and Grace, 2005). An important question regards the integration of prediction errors over time, to produce such pattern separation: since the second extinction trial in the gradual extinction condition was not reinforced, one could argue that at that point the rat experienced a similar prediction error to the one experienced in the first trial in standard extinction. To explain why a new memory was not formed in the case of gradual extinction, I posit that the pattern separation process determines similarity and infers latent causes by integrating prediction errors over a longer timescale than a single trial. From a functional perspective, such integration would be adaptive, preventing the animal from creating gratuitous memories in a noisy environment by smoothing the input signals over time.

Several previous studies have examined the effect of partial reinforcement during extinction. Bouton and colleagues (Bouton et al., 2004; Woods and Bouton, 2007) explored how occasional reinforced trials uniformly distributed throughout extinction training affect the rate of reacquisition after extinction training. Normally, animals show accelerated conditioning in a second training session after extinction was attenuated); this rapid reacquisition is consistent with a model in which the acquisition

memory is protected from disruption by extinction training, and can subsequently be retrieved to facilitate re-learning. Bouton and colleagues found that occasional reinforced trials led to slower reacquisition. This finding is consistent with my hypothesis that partial reinforcement during extinction prevents the splitting-off of a new memory; as a result, the original fear memory is modified during extinction, resulting in slower relearning. In the same set of experiments, Bouton et al. (2004) also found that a gradual extinction procedure (in which reinforcement was gradually removed) was effective at slowing reacquisition. This shows that the predictions of my model hold for appetitive as well as aversive conditioning. However, Bouton et als experiments differed from my own not only in terms of the valence of reinforcement, but also in that their gradual reductions were performed across sessions rather than within a session, and they used the speed of reacquisition to measure preservation of the original fear memory. Because a reacquisition test involves new learning, it is difficult to isolate from their experiments the effects of the procedure on the CS-US memory from effects on subsequent learning. Nonetheless, these results are consistent with my theoretical account of gradual extinction.

Using the rabbit nictitating membrane preparation, Kehoe and White (2002) showed that gradual reductions in unconditional stimulus intensity produced proportional reductions in the conditioned response. However, they found between-session spontaneous recovery, indicating that their procedure was ineffective at persistent attenuation of the conditioned response. Although their procedure differs in many details from the one described here, one important difference that might have led to spontaneous recovery in the experiments of Kehoe and White is that they used a reduction in intensity, rather than frequency, of the unconditional stimulus. If the subjective perception of aversive stimuli is not linear in their intensity, gradual reductions in intensity may still result in the experience of an abrupt change. This would generate a segmentation signal and lead to formation of a new memory trace.

Indeed, unpublished pilot experiments from Marie Monfils' laboratory suggest that intensity reduction is generally less effective at attenuating spontaneous recovery than frequency reduction.

In summary, my experimental results demonstrate the paradoxical effect that more tone-shock pairs can result in reduced return of fear, in line with my theoretical predictions. I interpret these results as showing that gradually reducing the frequency of tone-shock pairs leads to gradual modification of the original memory. In contrast, gradually increasing an initially low frequency does not attenuate recovery as a new memory is already formed early in extinction. My results provide support for recent theories of associative learning that are based on the interplay between error-driven learning and memory formation processes, and might suggest avenues for clinical treatment of disorders characterized by persistence of fear memories such as phobias and post-traumatic stress disorder.

Chapter 6

Statistical computations underlying the dynamics of memory

In the last chapter, I described the results of Pavlovian fear conditioning experiments which suggest that gradual changes in observational statistics favors the creation of a single memory trace, whereas abrupt changes favor the creation of multiple traces. In this chapter, I describe a visual memory task in humans which suggests a similar conclusion. I describe a variant of the latent cause framework that is more explicitly focused on change detection. The theoretical question is this: if the brain is confronted with a continuous stream of experience, where does one trace end and the next begin? Theorists have offered radically different answers to this question. According to biologically inspired theories (e.g., Hopfield, 1982; McClelland and Rumelhart, 1985; McNaughton and Morris, 1987), input patterns are assimilated into a distributed network of interconnected neurons. When allowed to run freely or with partial input, this network will converge to one or more stable configurations—*attractors*—corresponding to blends of stored input patterns. This view of memory asserts that

experiences are not stored individually, but rather overlaid on one another. Many modern psychological theories (e.g., Raaijmakers and Shiffrin, 1981; Hintzman, 1988; Nosofsky, 1988) adopt a diametrically opposed view: each input pattern is stored separately, and memory blending occurs at retrieval, rather than during storage.

One way to approach these issues is to consider the information processing problem being solved by the memory system. If we were to design a brain, what kind of memory traces would we want it to store? This exercise in "rational analysis" (Anderson, 1990) leads us to a statistical formulation of the memory storage problem. Building on the models described in chapters 3 and 4, I propose that the memory system is designed to facilitate optimal predictions under a particular generative model of the environment. According to this generative model (see also Yu and Dayan, 2005; Daw and Courville, 2008), the environment changes slowly over time, with occasional jumps between different "modes." Stored memories correspond precisely to inferences about the latent modes: input patterns are clustered together into a common memory trace if they are inferred to have been generated by the same mode. This theory retains the idea from the psychology literature that the memory system contains multiple traces, but assumes that each trace may be a blend of several input patterns, as is the case for many neural network models.

I show how this theory can illuminate several behavioral and neural phenomena relating to the dynamics of memory trace formation. I then describe a new behavioral experiment in which I present dynamically changing visual stimuli to subjects, and subsequently ask them to reconstruct one of the stimuli from memory. When the stimuli change gradually, subjects behave as though they formed one memory trace that adapts over time; when the stimuli change abruptly, subjects behave as though they formed two memory traces, one before the change and one after. My theory provides a good fit to the data, suggesting that statistical computations underlie the dynamical nature of memory traces.

6.1 Background: psychophysics and neurophysiology

Recent psychophysical studies have explored the dynamics of memory updating by presenting subjects with sequences of visual stimuli and then probing their ability to discriminate between different stimuli in the sequence. The logic of these studies is that if the stimuli are assimilated into the same memory trace, then discrimination will be poor; alternatively, if the stimuli are segmented into separate traces, discrimination will be good. For example, Wallis and Bülthoff (2001) presented subjects with a rotating face that gradually morphed into a different face. Compared to a condition in which the morphs were presented in a mixed (scrambled) order, participants in the gradual morph condition were more prone to perceive the different faces as belonging to the same person as the original face. Similar findings were reported by Preminger and colleagues (Preminger et al., 2007, 2009) using a variety of memory tests.

These psychophysical observations are complemented by neurophysiological studies of spatial representation in the rodent hippocampus. Many neurons in the CA3 subfield respond selectively when the animal is in a particular region of space, and are therefore known as "place cells." Large changes to the environmental context result in "global remapping" (a complete reconfiguration of the place fields), while small changes result in "rate remapping" (changes in firing rate while maintaining the same place fields) (Colgin et al., 2008). We can apply the same logic used in the psychophysical studies described above to the hippocampal representation of space: when one environment is morphed into another, will we see rate remapping (indicating a gradually changing memory) or global remapping (indicating the formation of a new memory)? Wills et al. (2005) had rats explore a set of boxes whose shape varied between square and circle (including intermediate shapes). They found that most place cells had different place fields for circle and square boxes, and these cells tended

to abruptly switch fields at the same intermediate shape. Another study using morphing boxes (Leutgeb et al., 2005) found a quite different result: place cells appeared to gradually change their fields from one environment to another. A crucial difference between these studies was that in the former a scrambled order of morphs was used, whereas in the latter the morphs were presented in consecutive order. Thus, the differences between these experimental results is consistent with the psychophysical studies described above, emphasizing the importance of sequential structure in the formation of memories.

Using a Hopfield network to encode the input patterns, Blumenfeld et al. (2006) proposed a "salience-weighted" modification of the standard Hebbian learning rule to model the experimental findings described above. Intuitively, the salience weight encodes a prediction error (or novelty) signal that indicates the extent to which none of the network's existing attractors match the input pattern.[1] A large salience weight promotes the formation of a new attractor based on the current input. For present purposes, the key idea to take away from this model is that prediction errors are useful signals for determining when to create new memory traces. In the network explored by Blumenfeld et al., a new attractor is only formed if the prediction error is sufficiently large, but how large is "sufficient"? In the next section, I place these ideas within a statistical framework, allowing us to specify the prediction error threshold in terms of probabilistic hypotheses about the environment.

6.2 The statistical framework

The essence of my approach is captured by the following generic assumption about the world: properties of the world usually change gradually, but occasionally undergo "jumps" that reflect a new underlying state of affairs. For example, when you walk

[1]Formally, the salience weight is the Hamming distance between the input pattern and the network state after one step of dynamics. The salience weight is updated incrementally after each input pattern so as to smooth across recent history.

around outside, you may experience gradual changes in temperature over the course of the day. If you step into a building, the temperature may jump abruptly. In predicting what the temperature will be like in 5 minutes, we might generalize from one outdoor location to another, but not between the indoor location and outdoor locations. Thus, our generalizations depends strongly on how we segment our observations; cognitively speaking, these segmentations are encoded in memory traces that aggregate observations assigned to the same segment.

The problem of estimating the current state of a hidden variable given previous sensory measurements is known in engineering as *filtering*. The classic example of a filtering algorithm is the Kalman filter (Kalman, 1960), which is the Bayes-optimal estimator under the assumption that the environment evolves according to a linear-Gaussian dynamical system (LDS). One way to accommodate jumps is to posit a collection of different dynamical modes (each corresponding to an LDS), and allow the generative process to switch between them stochastically. This is known as a switching LDS, and its corresponding Bayes-optimal estimator is the switching KF. To deal with real-world sensory measurements, it is not practical to specify in advance a finite number of modes; I therefore adopt a Bayesian nonparametric (infinite-capacity) generalization of the switching LDS based on the Dirichlet process (Fox et al., 2011), which allows the number of modes to adaptively expand as more measurements are collected. This model belongs to the latent cause framework introduced in Chapters 3 and 4; the dynamical modes can be thought of as latent causes that parameterize not only the observations but also the change process itself.

6.2.1 Generative model

Here I describe the generative model formally. Let $\mathbf{s}_t \in \mathbb{R}^D$ denote a set of sensory measurements at time t, arising from a hidden state variable $\mathbf{x}_t \in \mathbb{R}^D$. Let $z_t \in \{1, \ldots, \infty\}$ denote a dynamical mode (i.e., a cluster that specifies a particular

state-space dynamics). My model assumes that measurements are generated by the following stochastic process. For each time point t:

1. Draw a mode z_t from the CRP (Aldous, 1985) introduced in Chapter 2:

$$p(z_t = k|\mathbf{z}_{1:t-1}) \propto \begin{cases} N_k & \text{if } k \text{ is a previously sampled mode} \\ \alpha & \text{if } k \text{ is a new mode,} \end{cases} \tag{6.1}$$

   where N_k is the number of previous timepoints assigned to mode k, and $\alpha \geq 0$ is a concentration parameter. Eq. 6.1 corresponds to the distribution over partitions induced by the Dirichlet process.

2. If z_t is a new mode, draw the state variable $\mathbf{x}_t$ from a Gaussian base measure: $\mathbf{x}_t \sim \mathcal{N}(\boldsymbol{\mu}_0, c\mathbf{I})$, where c is the sensory noise variance. Otherwise, drift the state variable from its value when mode z_t was last active (indexed by τ): $\mathbf{x}_t \sim \mathcal{N}(\mathbf{x}_\tau, q\mathbf{I})$, where q is the drift variance. Note that the state variable for a mode is "frozen" when that mode is inactive.

3. Emit the sensory measurements by corrupting the state variable with Gaussian noise: $\mathbf{s}_t \sim \mathcal{N}(\mathbf{x}_t, r\mathbf{I})$.

This generative model is a simplification of the nonparametric switching LDS described in Fox et al. (2011). When $\alpha = 0$, the probability of a jump is 0 and we obtain a special case of the standard LDS formulation.

6.2.2 Bayesian inference

The Bayesian filtering problem is to infer the posterior distribution over the state variable $\mathbf{x}_t$ given the history of sensory measurements $\mathbf{S}_{1:t} = \{\mathbf{s}_1, \ldots, \mathbf{s}_t\}$. According

to Bayes' rule,

$$\begin{aligned} p(\mathbf{x}_t|\mathbf{S}_{1:t}) &\propto p(\mathbf{s}_t|\mathbf{x}_t) \sum_{\mathbf{z}_{1:t}} p(\mathbf{x}_t|\mathbf{S}_{1:t-1}, \mathbf{z}_{1:t}) p(\mathbf{z}_{1:t}) \\ &\approx p(\mathbf{s}_t|\mathbf{x}_t) \sum_{z_t} p(\mathbf{x}_t|\mathbf{S}_{1:t-1}, z_t) p(z_t|\hat{\mathbf{z}}_{1:t-1}), \end{aligned} \tag{6.2}$$

where $\hat{z}_t = \arg\max_k p(z_t = k|\mathbf{S}_{1:t}, \hat{\mathbf{z}}_{1:t-1})$. This corresponds to a simple "local" *maximum a posteriori* approximation (Anderson, 1991; Sanborn et al., 2010; Wang and Dunson, 2011) that maintains only a single partition, $\hat{\mathbf{z}}_{1:t}$.[2] The posterior over mode assignments is given by:

$$p(z_t = k|\mathbf{S}_{1:t}, \hat{\mathbf{z}}_{1:t-1}) \propto p(\mathbf{s}_t|\hat{\mathbf{z}}_{1:t-1}, z_t = k) p(z_t = k|\hat{\mathbf{z}}_{1:t-1}), \tag{6.3}$$

where the first term is the likelihood:

$$p(\mathbf{s}_t|\hat{\mathbf{z}}_{1:t-1}, z_t = k) = \begin{cases} \mathcal{N}(\mathbf{s}_t; \mathbf{x}_\tau, \Sigma_\tau + (r+q)\mathbf{I}) & \text{if } k \text{ is a previously sampled mode} \\ \mathcal{N}(\mathbf{s}_t; \mathbf{0}, (r+c)\mathbf{I}) & \text{if } k \text{ is a new mode.} \end{cases} \tag{6.4}$$

The second term in Eq. 6.3 is the prior (Eq. 6.1). I also used the local MAP approximation in Chapter 4.

The conditional distribution $p(\mathbf{x}_t|\mathbf{S}_{1:t-1}, z_t)$ is Gaussian with mean $\hat{\mathbf{x}}_t$ and covariance Σ_t:

$$\hat{\mathbf{x}}_t = \hat{\mathbf{x}}_\tau + \mathbf{K}_t(\mathbf{s}_t - \hat{\mathbf{x}}_\tau), \qquad \Sigma_t = (\mathbf{I} - \mathbf{K}_t)\Sigma_\tau, \tag{6.5}$$

where $\mathbf{K}_t = (\Sigma_\tau + q\mathbf{I})[\Sigma_\tau + (r+q)\mathbf{I}]^{-1}$ is known as the *Kalman gain*. Note that τ is

[2] Although I could have used more sophisticated methods (e.g., particle filtering) to approximate the marginalization, this method works equally well on the examples I consider, and is much faster (making it easier to fit to behavioral data, as described below).

implicitly a function of the mode assignments, $\hat{\mathbf{z}}_{1:t-1}$. This completes the description of my inference algorithm, which I refer to as the *Dirichlet process Kalman filter* (DP-KF).

6.2.3 Illustrations

Eq. 6.3 operationalizes the idea that large prediction errors will lead to memory trace formation: when $||\mathbf{s}_t - \mathbf{x}_\tau||$ is large relative to $||\mathbf{s}_t||$, the DP-KF will tend to assign observation t to a new mode, analogous to the process by which the Blumenfeld et al. (2006) saliency-weighted learning rule creates a new attractor when the input pattern fails to match any of the existing attractors. Figure 6.1 (left) illustrates this process. The sensory measurements drift gradually, undergo a jump, and then drift gradually again. The standard KF (squares) smooths across the jump, whereas the DP-KF (circles) tries to find piecewise smoothness by segmenting the time series into two modes, thereby producing better predictions.

The right panel of Figure 6.1 shows the results of applying the DP-KF to the "gradual" and "mixed" protocols described in Section 6.1. I used a sequence of one-dimensional measurements morphing between 0 and 1. In the gradual protocol, the morph index increases monotonically with time, whereas in the mixed protocol the morphs are presented in scrambled order. To analyze the simulated data, I re-sorted the indices from the mixed condition to match the gradual condition and calculated the posterior probability of mode 1. Consistent with the psychophysical and neurophysiological data (Wallis and Bülthoff, 2001; Preminger et al., 2007, 2009; Wills et al., 2005; Leutgeb et al., 2005), we see that the mixed protocol results in morphs being assigned to two different modes, whereas the gradual protocol results in all the morphs being assigned to the same mode.

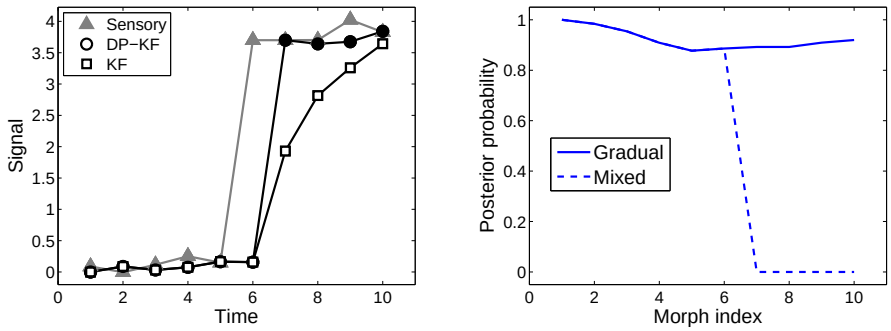

Figure 6.1: **Simulations**. (*Left*) Simulated sensory measurements and inferred state variables. For the DP-KF, the colors indicate the mode assignment with the highest posterior probability, white = mode 1, black = mode 2. Model predictions are computed *before* the sensory measurement. (*Right*) Posterior probability of mode 1 as a function of morph index in the gradual and mixed protocols, using the DP-KF. See text for details.

6.3 Experiment: reconstruction of dynamically changing visual stimuli

In this section, I describe an experiment designed to test a basic prediction of my model. Figure 6.2 shows the interface for the task. I exposed subjects to sequences of simple visual stimuli (oriented lines of varying lengths), asking them on each trial to predict the orientation and length of the next line. In the "gradual" condition, the lines diffused through orientation/length space; in the "jump" condition, the diffusion was interrupted by an abrupt jump in the middle of the sequence. At the end of each sequence, subjects were asked to reconstruct one of the lines from the beginning of the sequence.

I reasoned that if subjects use prediction errors to segment their observations into distinct memory traces, then they would create two traces for the jump condition (one for the first half and one for the second half of the sequence), but only one trace for the gradual condition. If subjects segment the sequence, then memory for the first half should be unaffected by observations in the second half. I therefore hypothesized that reconstructions of early lines would be relatively veridical in the jump condition.

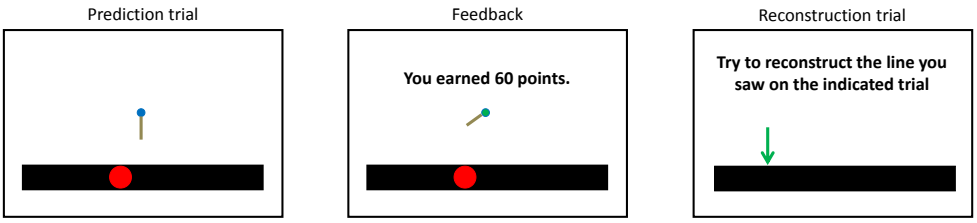

Figure 6.2: **Experimental task**. (*Left*) Prediction trial: subjects were asked to predict the orientation and length of a line segment (shown in the center of the screen). After making their predictions, they were shown the true line segment and a point score based on their accuracy. At the bottom of the screen, a red circle superimposed on a timeline (the black bar) indicates the trial's serial position in the block. (*Middle*) After making a prediction, subjects were shown the true line and awarded points in proportion to their accuracy. (*Right*) Reconstruction trial: at the end of each block, participants were asked to reconstruct the line they saw on one of the first three trials (indicated by an arrow).

By contrast, in the gradual condition, later observations will be assigned to the initial mode, leading to alteration of the memory trace. Compared to the jump condition, reconstructions in the gradual condition should therefore be more similar to later lines and less similar to the early lines.

6.3.1 Methods

Subjects. 32 undergraduates received course credit or payment for participating in the experiment. The experiment was approved by the Institutional Review Board.

Stimuli. The stimuli consisted of oriented line segments that diffused gradually through orientation/length space. In generating trajectories through this space, I required the Euclidean distance between the start and end points to lie within a narrow range (60-70 percent of the maximum). I also constrained the diffusion to always move at a 45 degree angle through the space (i.e., there was always an equal amount of change in both dimensions), so as to encourage subjects not to attend differentially to one dimension. Jumps were also at a 45 degree angle, but traversed a distance 4 times as long as the other steps. Jumps always occurred in the middle of the trajectory and were unsignaled. Examples of jump and gradual trajectories are

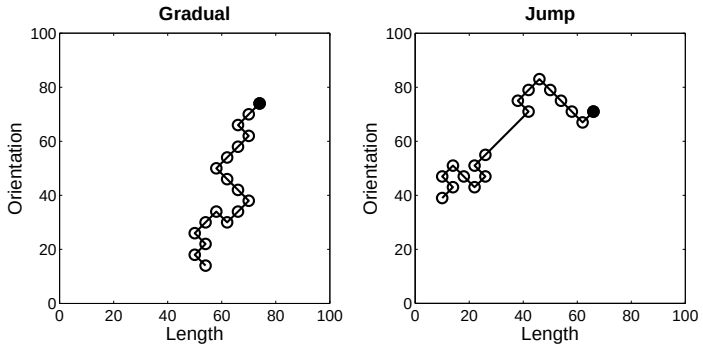

Figure 6.3: **Example trajectories**. Each circle represents a line segment presented to subjects in a sequence, with the shaded circle indicating the starting point. The dimensions are standardized to a [0,100] range. (*Left*) A "gradual" trajectory. (*Right*) A "jump" trajectory.

shown in Figure 6.3.

Procedure. Subjects played 12 blocks of the task (half of which were jump trajectories). Each block consisted of a sequence of 18 line segments. A timeline showed subjects the serial position of each trial in a block. On each trial, subjects were asked to adjust the orientation and length of a line on the screen so as to predict the next observed line. After making their prediction, subjects were shown the true line and awarded points based on how accurate their prediction was. At the end of the block, subjects were shown an arrow pointing toward a point on the timeline and asked to reconstruct the line they saw on that trial. Subjects were always asked to reconstruct one of the first 3 lines. No feedback was given on reconstruction trials.

Model-fitting. The noise variance r, the dynamics variance q and the concentration parameter α were treated as free parameters and fit to each subject's data separately by minimizing the sum squared difference between model predictions and human predictions. For the KF model, α was set to 0; thus, this model has one fewer free parameter. To approximate an uninformative prior over $\mathbf{x}_0$, I fixed $c = 1000$. I fixed $\boldsymbol{\mu}_0 = [50, 50]$, since that is the most agnostic prior in my paradigm. The models were not fit to the reconstruction trials.

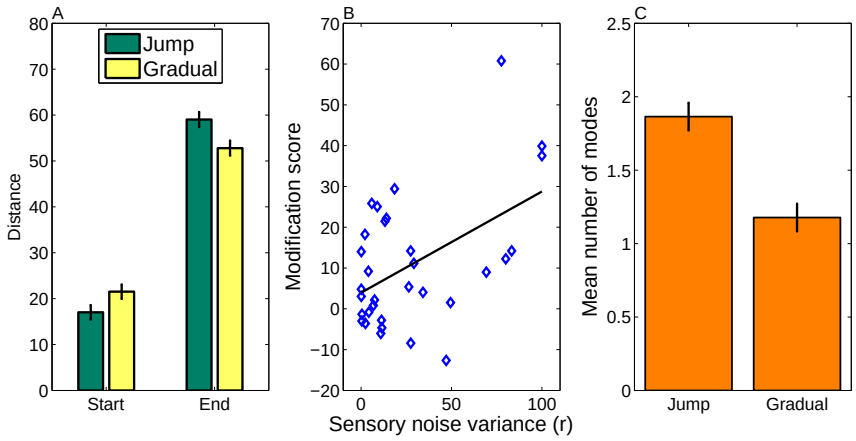

Figure 6.4: **Experimental results**. (A) Euclidean distance between subjects' reconstructions and the first and last lines in a block. (B) Modification score (see text for details) plotted against each subjects' fitted sensory noise variance parameter. (C) Number of clusters inferred by the DP-KF model for each condition.

6.3.2 Results

I fit the DP-KF and KF models to the responses on prediction trials, holding out the reconstruction trials for validation. I then compared participants' reconstructions to simulations of the fitted DP-KF model. I calculated the Euclidean distance between subjects' reconstructions and the true line at the beginning of the block, as well as the distance from the true line at the end of each block (Figure 6.4A). As predicted, subjects' reconstructions were closer to the first line in the jump condition than in the gradual condition ($t = 2.88, p < 0.01$), and this pattern reversed for distance to the last line ($t = 3.86, p < 0.001$). A two-way (first/last $\times$ gradual/jump) ANOVA confirmed that the interaction is significant ($F = 14.42, p < 0.001$). I interpret this result as follows: subjects formed one memory trace in the gradual condition, causing lines from the second half to influence memory for the lines from the first half. In the jump condition, subjects formed two memories (one for each half), effectively protecting memory of the lines shown in the first half from distortion by the lines from the second half of the block.

I then compared subjects' reconstruction data to parameter estimates and predic-

tions from the fitted DP-KF model. Note that reconstruction data were not included when fitting the model, so these analyses constitute independent tests of the model. I found that the "modification score" (defined as the subject-specific interaction contrast used in the ANOVA[3]) correlated with parameter estimates of r, the sensory noise variance (Figure 6.4B; $r = 0.49, p < 0.005$). A higher modification score means that a subject shifted their reconstructions away from the beginning of the block and closer to the end of the block. Intuitively, if a subject expects more noise in the environment, then she will also be less likely to infer a new mode after the jump, which (in turn) should result in increased modification of the original memory. Further supporting my model-based interpretation, I found that the number of dynamical modes inferred by the fitted DP-KF model was, on average, 1.89 in the jump condition, and 1.2 in the gradual condition (Figure 6.4C). The difference between these two conditions was significant ($t = 6.95, p < 0.001$).

Figure 6.5 compares the DP-KF reconstructions against the human reconstructions (aggregating across all subjects and conditions). It is evident that the model is well matched to human behavior. To assess this relationship statistically, I computed the Pearson correlation coefficient for each subject separately, Fisher z-transformed this value, and performed a t-test against 0. Separate correlations for orientation and length were both significant ($p < 0.0001$).

I performed a quantitative comparison between the DP-KF and KF models in two ways. First, I approximated the Bayes factor between the two models using the Bayesian information criterion (Kass and Raftery, 1995). This measure (which penalizes model complexity) strongly supported the DP-KF model (Figure 6.6, left): 28 out of 32 subjects had a log Bayes factor greater than zero. The median log Bayes factor was significantly greater than 0 according to a Wilcoxon signed-rank test ($p <$

[3]Let d_t^j be the distance between the reconstruction and the ground truth line on trial t in the jump condition. Let d_t^g denote the same distance for the gradual condition. Then the modification score is $\langle d_1^g - d_1^j + d_{18}^j - d_{18}^g \rangle$, where the average is taken over blocks.

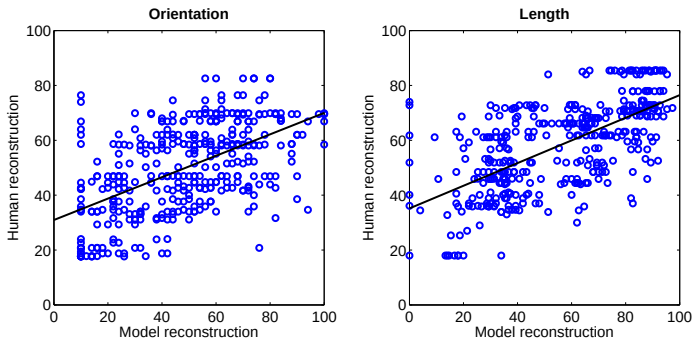

Figure 6.5: **Comparison of model and human reconstructions**. Each point corresponds to one reconstruction for one subject. Length and orientation reconstructions are plotted separately, though subjects make both simultaneously. Least-squares lines are superimposed on the data.

0.001). Second, I calculated the squared error between the models' reconstructions and human reconstructions. This analysis also supported the DP-KF model (Figure 6.6, right): the median squared error difference was significantly less than 0 according to a signed-rank test ($p < 0.001$).

6.4 Discussion

In this chapter, I addressed, both theoretically and experimentally, a basic question about memory: when are old traces modified, and when are new traces formed? My answer took the form of a rational analysis. In particular, I proposed that memories traces arise through a process of optimal filtering in a dynamically changing environment. New traces are formed when there are abrupt discontinuities in the temporal dynamics of the environment. Such discontinuities are typically accompanied by a large prediction error, suggesting a biologically plausible mechanism for implementing trace formation. Prediction errors are known to be computed in many areas of the brain, including area CA1 of the hippocampus (Vinogradova, 2001) and the midbrain dopaminergic nuclei (Bayer and Glimcher, 2005). Indeed, predictive coding theories propose that prediction errors are computed throughout the neocortex (Friston,

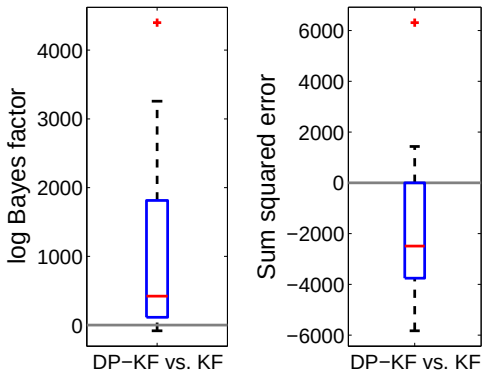

Figure 6.6: **Model comparison**. (*Left*) Log Bayes factors comparing the DP-KF and KF models of human behavior on the prediction trials. Values greater than 0 support the DP-KF model. (*Right*) Relative sum squared error for DP-KF vs. KF models. Errors are calculated as the squared difference between model and human reconstructions. Values less than 0 support the DP-KF model.

2005).

Several authors have proposed neural implementations of the KF (Denève et al., 2007; Wilson and Finkel, 2009). Wilson and Finkel (2009) derived an approximation of the KF that can be computed by a recurrent neural network when the prediction error is small. Intriguingly, when the prediction error is large, their approximation breaks down by creating two bumps in the posterior (rather than one as in the exact KF) with each bump implementing an independent KF. My theory suggests a normative account of this feature, since a network that creates multiple bumps is precisely what is required by the DP-KF algorithm. Pursuing this connection is an exciting direction for future research.

Work on change detection (e.g., Nassar et al., 2010; Summerfield et al., 2011) addresses a similar question: how does the brain detect a change in the statistics of sensory signals? The study of Nassar et al. (2010), for example, showed that humans use the recent history of prediction errors to determine when a change has occurred. This work differs from my own in several ways. First, most existing change detection theories assume that the sensory statistics are stationary between jumps, whereas I

allow for gradual change between jumps. Second, once a jump has occurred, theories of change detection assume that the statistics of earlier epochs are discarded forever; in contrast, my model assumes that the environment can return to earlier modes. This is particularly important with regard to my account of the line reconstruction task, which requires subjects to recall earlier observations. If subjects discarded the statistics associated with stimuli before a jump, we would expect them to perform poorly on the reconstruction task (which was not the case).

6.4.1 Conclusions

In this chapter, I investigated empirically a fundamental prediction that change detection models make for memory. If, as I hypothesize, new experience is incorporated into old memory traces based on similarity, then abrupt change (i.e., dissimilar data) will prompt the creation of a new memory trace, thereby protecting old memories from being modified by new data, whereas gradual change will not. To my knowledge, my work is the first to make this prediction for memory processes and the first to test this prediction empirically in humans.

Chapter 7

Occam's razor in categorical perception

In this chapter, I explore the consequences of a particular inductive bias—simplicity—implied by the Chinese restaurant process prior employed by the models in chapters 3, 4 and 6. The model presented later in this chapter is another variant of the latent cause framework, but applied to a visual judgment task. Together with the visual memory task described in the last chapter, the experiments and simulations in this chapter demonstrate the generality of the latent cause framework's predictions.

The 14th-century English friar and theologian William of Occam advised philosophers "not to multiply entities beyond necessity" (Boehner, 1957). The contemporary interpretation of Occam's razor is that, all other things being equal, simpler explanations of data should be preferred to more complex explanations. This heuristic notion has found mathematical expression in Bayesian statistics (Jaynes, 2003) and algorithmic information theory (Li and Vitanyi, 2008). It has since been applied to cognitive psychology as the "simplicity principle" (Chater and Vitányi, 2003; Feldman, 2003): the idea that humans seek simple explanations of their sensory input. My focus in this paper is on unsupervised category learning, where evidence suggests

that humans assign stimuli to a small set of categories, only inventing new categories when the stimulus statistics change radically (Anderson, 1991; Clapper and Bower, 1994; Pothos and Chater, 2002; Love et al., 2004; Sanborn et al., 2010).

If the categories people invent dictate how they "carve nature at its joints" (i.e., divide the environment into meaningful entities; see Gershman et al., 2010), then effects of Occam's razor should be discernible in perceptual estimation. Substantial evidence exists that categories shape perception (Goldstone, 1995; Hemmer and Steyvers, 2009; Huttenlocher et al., 1991, 2000). For example, Goldstone (1995) showed that an object belonging to a shape category with typically red objects was judged to be more red than an identically colored object belonging to a different category. As another example, syllables belonging to different phonetic categories are more easily discriminated than syllables with the same physical difference but belonging to the same category—the so-called *perceptual magnet effect* (Liberman et al., 1957). However, these studies assume a given category structure, whereas many real-world learning situations (particularly during development) require one to discover the underlying category structure from undifferentiated sensory data. In these situations, I expect that Occam's razor will influence the number of perceptual categories inferred from sensory data, and in turn govern participants' estimates of stimulus properties. The experiments reported in this paper were designed to test this hypothesis.

The stimuli in my experiments consisted of randomly scattered colored circles displayed on the computer screen (Figure 7.1), similar to stimuli used in studies of number perception (Izard and Dehaene, 2008). Each trial was characterized by one of two colors, and all circles were displayed in this color. The number of circles on each trial was drawn from a color-specific Gaussian distribution. The distributions differed in their means (Experiments 1a and 1b) or variances (Experiment 2). Participants were asked to judge how many circles there were on the screen, but did not have enough time to count them explicitly.

If the distributions (corresponding to the two colors) overlap sufficiently, Occam's razor dictates that the stimuli should all be assigned to one category despite their obviously different colors, a prediction formalized in several models of categorization (Anderson, 1991; Sanborn et al., 2010). The consequence of merging the two perceptual categories is that estimates will be "regularized" towards the average of the two distributions. In contrast, reducing overlap between the distributions is expected to diminish this regularization, as it supports separate categories for each color. Each of the experiments reported below included a high overlap condition in which merging (and hence more regularization) was expected to occur, and a low overlap condition in which splitting (and less regularization) was expected to occur.

To make my theoretical account explicit and quantitative, in Section 7.4 I present a computational model of human performance in my task. In the spirit of the probabilistic motivation for Occam's razor described above, I derive my model from hypothesized probabilistic assumptions about the environment and suggest that participants perform approximately optimal inference. In other words, I undertake a "rational analysis" (Anderson, 1990). My aim is to elucidate the computational constraints, rather than particular processing or implementational mechanisms, that govern categorical perception in my task.

7.1 Experiment 1a

In my first experiment, I manipulated categorical overlap by varying (within-subject) the distance between the means of the two distributions. One distribution (mean 65, standard deviation 10) was designated the "baseline" and did not change across blocks. On each block the second, "alternative" distribution either had low overlap (mean 35, standard deviation 10) or high overlap (mean 55, standard deviation 10) with the baseline distribution (Figure 7.2, left). I refer to these conditions as *Low*

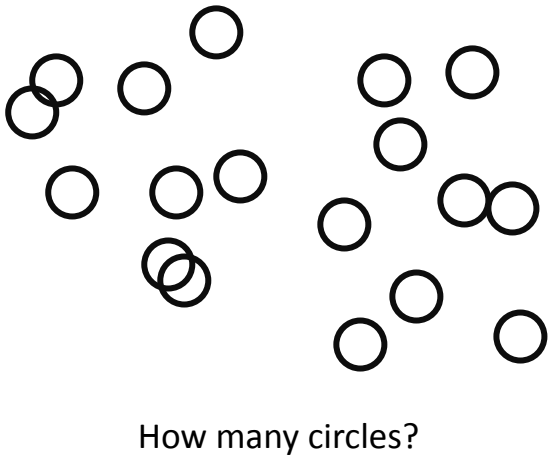

Figure 7.1: **Example trial**. On each trial, participants were presented with a random scattering of circles and asked to estimate the number of circles. The circles on each trial were all of the same color (color not shown here). The number of circles on each trial was drawn from a color-specific Gaussian distribution.

mean alternative and *High mean alternative*, respectively. Each distribution (alternative and baseline) was associated with a unique color. On each trial, the number of circles on the screen was randomly drawn from the distribution associated with the circles' color on that trial.

I did not instruct participants explicitly about color in any way, but I expected them to use it as a cue for categorization. Moreover, I expected use of the color cue to depend on a combination of sensory evidence (i.e., the number of circles) and the simplicity bias towards fewer categories. When the numbers of circles in all trials are similar, I expected color to be effectively ignored, and trials to be categorized together. In this case, estimates about the number of circles should be affected by the statistics of both colors. Specifically, I predicted that estimates on the baseline trials would be lower on average in the High mean alternative condition than in the Low mean alternative condition, due to the regularization induced by merging the color categories together in the High mean alternative condition. Note that if participants ignored color on all blocks I would expect a different result: baseline estimates in the High mean alternative condition should be systematically higher

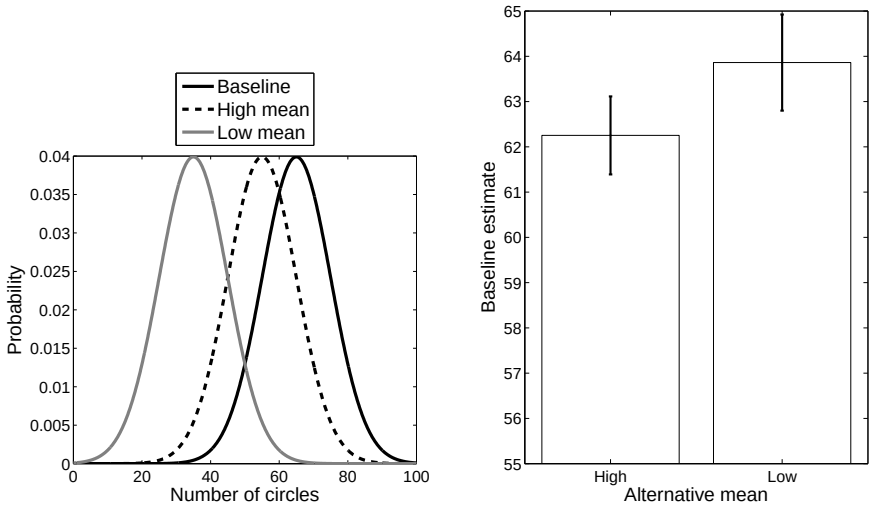

Figure 7.2: **Experiment 1a design and results**. (*Left*) Distributions for each category. (*Right*) Average estimates for the baseline category in each condition. Error bars represent standard error of the mean.

than in the Low mean alternative condition. Alternatively, if participants always used color as a categorization cue, there should be no difference between estimates of baseline trials in the two conditions, since the baseline distribution is the same in both cases.

7.1.1 Method

Participants

Fourteen students participated in the experiment for course credit or monetary compensation (10 dollars). All subjects gave informed consent and the study was approved by the Princeton University Institutional Review Board.

Procedure

Stimuli consisted of colored circles displayed in a random spatial configuration within a bounded section of the computer screen. On each trial, the participant was presented with a pattern of randomly scattered (occasionally overlapping) circles (Figure 1),

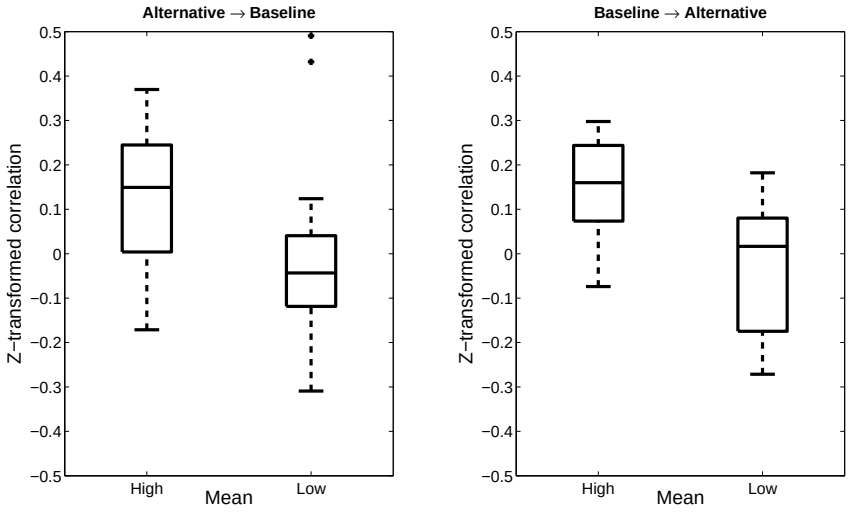

Figure 7.3: **Trial-wise correlations in Experiment 1a**. (*Left*) Fisher z-transformed correlations between estimates on baseline trials and on the preceding alternative trials. (*Right*) Correlations between alternative trials and the preceding baseline trials.

where the number of circles was drawn from a Gaussian with a category-specific mean and variance. There were two trial types: 'baseline' trials in which the number of circles was drawn from a Gaussian with mean 65 and standard deviation 10), and 'alternative' trials. In the 'High mean alternative' block the latter trials were drawn from a Gaussian with mean 55 and standard deviation 10. In the 'Low mean alternative' block, the alternative trials were drawn from a Gaussian with mean 35 and standard deviation 10. In all cases, the number of circles was truncated between 10 and 100, and rounded to the nearest integer. Each of the two categories in a block was randomly associated with a different color of circles (red, blue or green).

The participant was given 5 seconds to enter a 2-digit estimate of the number of circles on the screen using the keyboard; if no response was entered within this time limit, a message indicated that the response was too slow and the trial was subsequently not used in data analysis. After entering a response, the participant received feedback indicating the correct number of circles. Each subject performed 8 blocks of the High mean alternative condition and 8 blocks of the Low mean alterna-

tive condition (randomly interleaved), with 20 trials in each block (10 baseline and 10 alternative, randomly interleaved). All experiments were implemented in Matlab (Version 7.9.0.529) using the Psychophysics toolbox (Brainard, 1997).

7.1.2 Results and discussion

The average responses on baseline trials in each condition are shown in Figure 7.2 (right). Estimates of the number of circles on baseline trials in the High mean alternative condition (mean = 62.25) were significantly lower than in the Low mean alternative condition (mean = 63.86), $t_{13} = 2.41, p < 0.05$. This result is consistent with the hypothesis that participants are more likely to assign the alternative and baseline distributions to the same category in the High mean alternative condition (due to greater overlap) than in the Low mean alternative condition.

I also examined the estimates on alternative trials. The average number of circles reported by participants closely tracked the true average: 55.44 for the High mean alternative condition and 36.47 for the Low mean alternative condition. T-tests confirmed that average participant estimates were not significantly different from the true average ($p = 0.51$ for the High mean alternative condition and $p = 0.07$ for the Low mean alternative condition).

If participants are really merging baseline and alternative categories in the High mean alternative condition, one might argue that we should also see regularization effects on the alternative trials. While I saw no evidence for such regularization in the trial-averaged data, it may be the case that regularization effects operate over timescales that are shorter than a whole block. To test this hypothesis, I calculated the correlation between estimates on each baseline trial and the preceding alternative trial (note that, due to the randomized trial order, the preceding alternative trial might be several trials back). I reasoned that if estimates are influenced by recently experienced trials, then my correlation dependent measure should be positive. Impor-

tantly, this should only occur if both trials are assigned to the same merged category. Figure 7.3 (left) shows the results of this analysis: Fisher z-transformed correlations were significantly greater than 0 in the High mean alternative condition ($p < 0.01$, one-sample t-test) but not in the Low mean alternative condition ($p = 0.86$). I also examined the influence of baseline trials on subsequent alternative trials (Figure 7.3, right): Again, Fisher z-transformed correlations were significantly greater than 0 in the High mean alternative condition ($p < 0.001$, one-sample t-test) but not in the Low mean alternative condition ($p = 0.50$). These results are consistent with the hypothesis that the High mean alternative condition promotes category merging while the Low mean alternative condition does not.

The correlation analyses reported above also rule out an alternative explanation of my findings in terms of contrast effects. According to this explanation (see Holland and Lockhead, 1968), contrast between the baseline and alternative categories is accentuated in the Low mean alternative condition, causing participants to produce higher estimates for baseline trials compared to estimates in the High mean alternative condition. Such a contrast explanation would predict *negative* correlations between estimates in the baseline and alternative trial types; yet I found no evidence for negative correlations.

7.2 Experiment 1b

Experiment 1b was designed to replicate and extend the findings of Experiment 1a. For both the High mean alternative and Low mean alternative conditions in Experiment 1a, the alternative mean was lower than the baseline mean. In Experiment 1b, I examined whether the same effects would be found when the alternative means were higher than the baseline mean. Here I predicted that participants would be more likely to merge the baseline and alternative categories in the Low mean alternative

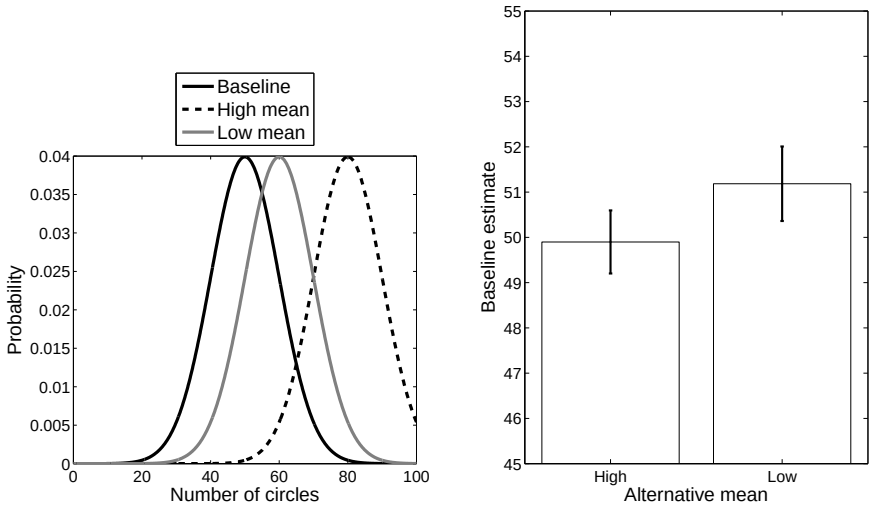

Figure 7.4: **Experiment 1b design and results**. (*Left*) Distributions for each category. (*Right*) Average estimates for the baseline category in each condition. Error bars represent standard error of the mean.

condition than in the High mean alternative condition; accordingly, baseline estimates should be regularized upward to a greater extent in the Low mean alternative condition.

7.2.1 Method

Participants

Fourteen students participated in the experiment for monetary compensation (10 dollars). All subjects gave informed consent and the study was approved by the Princeton University Institutional Review Board.

Procedure

The procedure in this experiment was identical to the procedure used in Experiment 1a, with only the category means changed. Specifically, I used the following category means: 50 for the baseline trials, 60 for alternative trials in the Low mean alternative condition, and 80 for alternative trials in the High mean alternative condition (see

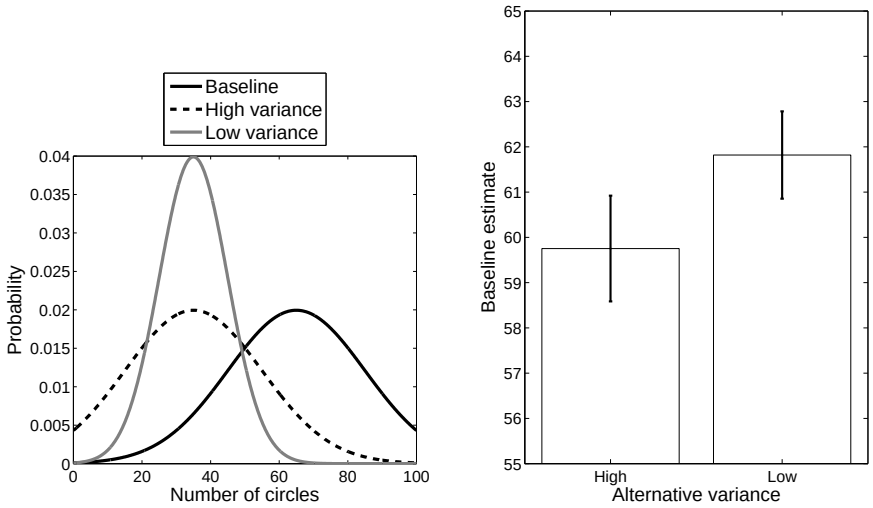

Figure 7.5: **Experiment 2 design and results**. (*Left*) Distributions for each category. (*Right*) Average estimates for the baseline category in each condition. Error bars represent standard error of the mean.

Figure 7.4, left).

7.2.2 Results and discussion

The average responses on baseline trials in each condition are shown in Figure 7.4 (right). Estimates of the number of circles on baseline trials in the High mean alternative condition (mean = 49.90) were significantly lower than in the Low mean alternative condition (mean = 51.18), $t_{13} = 2.36, p < 0.05$. This result is consistent with the hypothesis that participants are more likely to merge the alternative and baseline distributions together in the Low mean alternative condition (due to greater distributional overlap) than in the High mean alternative condition.

7.3 Experiment 2

My second experiment was identical to Experiment 1 in all respects except that I manipulated the variances of the distributions rather than their means, as illustrated

in Figure 7.5 (left). This manipulation was again expected to affect the probability of splitting or merging perceptual categories. Specifically, the High variance condition creates more overlap between the alternative and baseline distributions compared to the Low variance condition, leading to the prediction that estimates of baseline trials in the High variance condition will be regularized downward more than in the Low variance condition.

7.3.1 Method

Participants

Fourteen students participated in the experiment for course credit or monetary compensation (10 dollars). All subjects gave informed consent and the study was approved by the Princeton University Institutional Review Board.

Procedure

The procedure was identical to Experiments 1a and 1b, except that the alternative trials differed in their standard deviations. Both High and Low variance alternative trials had a mean of 35; High variance trials had a standard deviation of 20, while Low variance trials had a standard deviation of 10. Baseline trials (same for both conditions) had a mean of 65 and a standard deviation of 20.

7.3.2 Results and discussion

The average responses on baseline trials in each condition are shown in Figure 7.5 (right). Judgments of the number of circles on baseline trials in the High variance condition (mean = 59.75) was significantly lower than in the Low variance condition (mean = 61.82), $t_{13} = 2.72, p < 0.05$. This result is consistent with the hypothesis that participants are more likely to merge the alternative and baseline distributions

together in the High variance condition (due to greater overlap) than in the Low variance condition.

I also examined the judgments on alternative trials. Unlike in Experiment 1a, the average number of circles reported by participants deviated from the true average in the direction of the baseline average: 39.64 for the High variance condition and 37.45 for the Low variance condition. T-tests confirmed that average participant estimates were significantly different from the true average ($p < 0.001$ for the High variance condition and $p < 0.01$ for the Low variance condition). From a theoretical perspective, these results can be explained by the idea that with a larger standard deviation, category merging is more likely for both High and Low alternative trials compared to the alternative trials in Experiment 1a. Furthermore, the deviation (difference between estimated and true number of circles) was greater for the High variance condition than for the Low variance condition ($t_{13} = 2.63, p < 0.05$), consistent with my hypothesis that category merging (and hence more regularization) is more likely to occur in the High variance condition.

7.4 A rational analysis

In this section, I frame my experimental results in terms of a Bayesian computational model of the estimation task. This model constitutes a "rational analysis" (Anderson, 1990)—a specification of how an ideal observer would perform in my task. Although I do not necessarily believe that humans are precisely implementing Bayesian inference,[1] this analysis allows us to explore rather subtle hypotheses about cognitive processes, as I describe below.

According to the Bayesian framework (described formally in the next section), the computational problem facing a participant is to infer the posterior distribution over

[1]Nor do I necessarily believe that there is a unique ideal observer, since different priors lead to different inferences, all of which are rational from a statistical point of view.

the number of circles x_t on trial t, given noisy sensory input y_t, the circle color c_t, and the history of past trials. For a complete mathematical specification, I make several assumptions about the data-generating process. In particular, both the circle color and number are assumed to be governed by a latent perceptual category z_t drawn from some unknown number of categories. Thus, according to my rational analysis, the participant must implicitly average over her uncertainty about the latent categories in making her estimates. Importantly, I do not impute to the participant a fixed set of categories; rather, both the number and properties of the categories are inferred by the participant from her observational data. The simplicity principle enters into this model via the prior over categories: All other things being equal, the model has a preference for a small number of categories.

7.4.1 Generative process

The starting point of my rational analysis is the specification of a joint distribution over all the variables (both latent and observed) involved in the experimental task. This joint distribution is sometimes known as a *generative model*, since it represents the participant's (putative) assumptions about the process by which the observations were generated. The generative model I assume is a *mixture model*, where the number of circles x_t is drawn from a Gaussian distribution associated with the perceptual category $z_t = k$ active on trial t (I will use z_t and k interchangeably below to indicate categories, with the former used when categories on different trials need to be distinguished). The distribution over x_t is parameterized by a category-specific mean μ_k and standard deviation σ_k. The observed number of circles y_t (the noisy sensory signal) is drawn from a Gaussian distribution with mean x_t and standard deviation σ_y. The circle color $c_t \in \{1, \ldots, C\}$ is drawn from a category-specific multinomial distribution specified by parameters θ_k. In my experiments, $C = 3$.

I assume that participants begin each block with a prior belief about the param-

eters of the task. Specifically, I assume a normal-inverse-gamma prior on (μ_k, σ_k^2):

$$P(\mu_k, \sigma_k^2) = \mathcal{N}(\mu_k; \mu_0, \sigma^2/\eta_0)\text{IG}(\sigma_k^2; a_0, b_0), \tag{7.1}$$

where $\text{IG}(\cdot; a_0, b_0)$ is the probability density function of the inverse gamma distribution (see Gelman et al., 2004). The multinomial parameters for the color feature are assumed to be drawn from a symmetric Dirichlet distribution with parameter λ.

To complete the generative model, I need to specify a prior distribution on the set of category assignments, $\mathbf{z}_{1:t} = \{z_1, \ldots, z_t\}$, which can be understood as a partition of the observations into latent categories. I want to impute to the participant a prior that is flexible enough to entertain an unbounded number of possible categories. For this purpose, I choose the CRP introduced in Chapter 2, a prior over an unbounded number of partitions. Formally, the CRP is given by:

$$P(z_t = k|\mathbf{z}_{1:t-1}) = \begin{cases} \frac{M_k}{t} & \text{if } k \text{ is an old category} \\ \frac{\alpha}{t} & \text{if } k \text{ is a new category} \end{cases} \tag{7.2}$$

where M_k is the number of trials generated by category k up to trial t. The value of α controls the prior belief about the number of categories. As $\alpha \to 0$, all trials will tend to be assigned to the same category; in contrast, as $\alpha \to \infty$, each trial will be assigned to a unique category (the latter limiting case is closely related to exemplar models, as will be described below). The Chinese restaurant prior was in fact independently discovered by Anderson (1991) in the development of his rational model of categorization, and since then has been used in a wide variety of psychological models (e.g., Gershman and Niv, 2010; Kemp et al., 2010; Sanborn et al., 2010).

7.4.2 Posterior inference

Two computational problems face the participant. The first is to infer the posterior distribution over latent perceptual categories given a set of observations. This is done by inverting the generative model using Bayes' rule. The second is to use this distribution to estimate the "true" number of circles on the current trial (x_t) given noisy sensory input (y_t). Note that in my experiments all uncertainty about y_t disappears after feedback (i.e., when x_t is observed). The posterior computations below reflect probabilistic beliefs after feedback is observed. In the Appendix, I describe how predictions are computed before feedback, which I use to predict participant behavior.

The posterior over categories is stipulated by Bayes' rule:

$$P(z_t|\mathbf{c}_{1:t}, \mathbf{x}_{1:t}) \propto \sum_{\mathbf{z}_{1:t-1}} P(c_t|\mathbf{z}_{1:t}, \mathbf{c}_{1:t-1}) P(x_t|\mathbf{z}_{1:t}, \mathbf{x}_{1:t-1}) P(z_t|\mathbf{z}_{1:t-1}). \tag{7.3}$$

Using the shorthand $k = z_t$ and $c = c_t$, the conditional distributions are given by:

$$\begin{aligned} P(c_t|\mathbf{z}_{1:t}, \mathbf{c}_{1:t-1}) &= \int_\theta P(c_t|\theta, \mathbf{z}_{1:t}, \mathbf{c}_{1:t-1}) d\theta \\ &= \frac{\lambda + N_{ck}}{C\lambda + M_k} \end{aligned} \tag{7.4}$$

$$\begin{aligned} P(x_t|\mathbf{z}_{1:t}, \mathbf{x}_{1:t-1}) &= \int_\mu \int_{\sigma^2} P(x_t|\mu, \sigma^2, \mathbf{z}_{1:t}, \mathbf{x}_{1:t-1}) d\mu \, d\sigma^2 \\ &= T_{2a_k}\left(\frac{x_t - \hat{\mu}_k}{\beta_k}\right) \end{aligned} \tag{7.5}$$

where $T_{2a_k}(x)$ denotes the student t-distribution with $2a_k$ degrees of freedom, and

$$\hat{\mu}_k = \frac{\eta_0 \mu_0 + M_k \bar{x}_k}{\eta_k}, \tag{7.6}$$

$$\eta_k = M_k + \eta_0, \tag{7.7}$$

$$a_k = M_k + \frac{a_0}{2}, \tag{7.8}$$

$$b_k = b_0 + \frac{1}{2} \sum_{i=1}^{t-1} \delta[z_i, k](x_i - \bar{x}_k)^2 + \frac{M_k \eta_0 (\mu_0 - \bar{x}_k)^2}{2\eta_k}, \tag{7.9}$$

$$\beta_k = \frac{b_k(1 + \eta_k)}{a_k \eta_k}. \tag{7.10}$$

Here $\delta[\cdot, \cdot] = 1$ if its arguments are equal, and 0 otherwise. N_{ck} is the number of times category k was presented in conjunction with color c and $\bar{x}_k$ is the average number of circles observed for category k. These equations were derived from standard properties of the conjugate-exponential family of probability distributions (Gelman et al., 2004).

Intuitively, Eq. 7.4 keeps track of counts: The posterior $P(c_t | \mathbf{z}_{1:t}, \mathbf{c}_{1:t-1})$ will tend to concentrate around the color that was observed most often in conjunction with z_t (conditional on a particular instantiation of $\mathbf{z}_{1:t}$). The parameter λ regularizes the posterior towards the uniform distribution, taking into account the observer's prior uncertainty about the relationship between categories and colors. Similarly, Eq. 7.5 keeps track of category averages: The posterior $P(x_t | z_t, \mathbf{z}_{1:t-1}, \mathbf{x}_{1:t-1})$ will tend to concentrate around the average number of circles observed in conjunction with z_t.

The Bayes-optimal estimator of the number of circles x_t given noisy sensory input y_t is the posterior mean:

$$\mathbb{E}[x_t | y_t, \mathbf{x}_{1:t-1}, \mathbf{c}_{1:t}] = \sum_{\mathbf{z}_{1:t}} \int_x x P(x_t = x, \mathbf{z}_{1:t} | y_t, \mathbf{x}_{1:t-1}, \mathbf{c}_{1:t}) dx. \tag{7.11}$$

The estimated number of circles follows a mixture of Gaussians, where the mean of each mixture component is a weighted combination of the category mean and the

sensory input.

Because the sum in Eq. 7.3 is intractable to compute exactly, I resort to approximation methods. In the Appendix, I describe a particle filter algorithm (Doucet et al., 2001) for approximating the posterior with a set of samples (see Chapter 2). While this algorithm can be understood as a provisional hypothesis about how humans might approximate Bayes' rule in this task, it should be emphasized that my data do not directly discriminate between this hypothesis and other types of approximations.

7.4.3 Model-fitting and comparison

We cannot know what sensory input (y_t) a participant is receiving on each trial, so I made the expedient choice (following Huttenlocher et al., 1991, 2000) of setting $y_t = x_t$, which should be true on average, assuming participants are not systematically biased. I recenter the data by subtracting the empirical mean (true number of circles on average) from all the perceptual estimates, and therefore use $\mu_0 = 0$. I set $\lambda = 1$ and $b_0 = 10$ (which sets the scale of σ_k^2), fitting the remaining parameters $(\alpha, a_0, \eta_0, \sigma_y)$ using a hill-climbing algorithm. Each participant's data were fit with a different set of parameters. The objective function was the mean-squared error between the particle filter predictions (Eq. 7.13) and participants' estimates. This is equivalent to the assumption that behavioral responses are normally-distributed around the model predictions; the parameter values minimizing the objective function are thus maximum likelihood estimates.

7.4.4 Model predictions

Figure 7.6 shows the fitted model predictions for the baseline color in Experiments 1a, 1b and 2. While not in perfect quantitative agreement with the behavioral data (Figures 7.2, 7.4 and 7.5, respectively), the model reproduces the observed qualitative pattern: The High mean alternative condition leads to more regularization than the

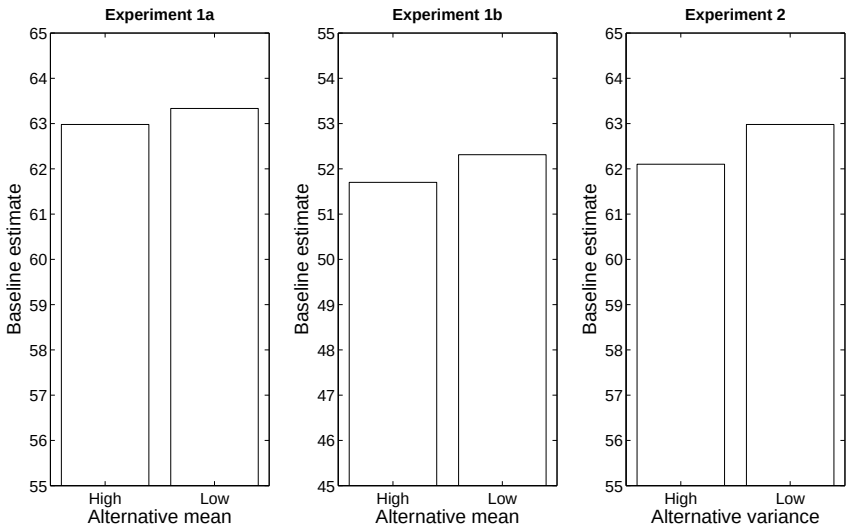

Figure 7.6: **Model predictions**. Estimates are derived from the fitted rational model for (*Left*) Experiment 1a, (*Middle*) Experiment 1b, and (*Right*) Experiment 2.

Low mean alternative condition (Experiment 1a), and the High variance condition leads to more regularization than the Low variance condition (Experiment 2). These effects arise in the rational model due to the fact that the greater overlap between the baseline and alternative distributions in the High mean/variance conditions increases the probability that trials with different-colored circles will be attributed to the same category (relative to the Low mean/variance conditions), thereby pushing estimates towards the aggregate mean of the two distributions.

7.4.5 Comparison to alternative models

The rational model I have been presented can be contrasted with a continuum of models that have been considered for perceptual estimation tasks. At one pole of the continuum is the model of Huttenlocher et al. (1991), which, in the context of my experiments, endows each color with its own category. The perceptual estimate on a given trial is assumed to be regularized towards the mean associated with the color on that trial (see also Hemmer and Steyvers, 2009; Huttenlocher et al., 2000). This model cannot explain my findings, since it predicts that regularization will always

be in the direction of the color-specific mean, disallowing perceptual categories that collapse across color (see Sailor and Antoine, 2005). In other words, the model of Huttenlocher et al. (1991) does not accommodate the possibility of adaptive category merging.

At the other pole is the family of exemplar models, which have proven successful in accounting for human categorization, identification and recognition memory (Kruschke, 1992; Medin and Schaffer, 1978; Nosofsky, 1986, 1988). The essential idea underlying these models is that estimates are formed by comparing the current stimulus to a stored set of memory traces (exemplars). As was recognized by Nosofsky (1991) in his discussion of Anderson's rational model of categorization (Anderson, 1991), the rational model becomes equivalent to the exemplar model in the limit $\alpha \to \infty$. In this limit, the number of categories inferred by the model is equal to the number of observations; hence, each category corresponds to an episodic memory trace, and Bayesian estimates correspond to averages of these traces in the same fashion as the exemplar model. In a sense, the exemplar model postulates the least parsimonious representation of the subject's perceptual inputs, since commonalities between observations are not explicitly abstracted.

It is difficult to rule out an exemplar explanation of my findings through examination of means in each condition. Instead, I undertook a quantitative model comparison to compare my model to the exemplar extreme. First, I compared the evidence for each model on a subject-by-subject basis. Model evidence was quantified by the Bayesian Information Criterion approximation to the Bayes Factor (Kass and Raftery, 1995), which balances fit to data against model complexity. Note that the rational model has one more parameter (α) than the exemplar model, and is therefore more complex. The model comparison analysis strongly supported the rational model over the exemplar model (Figure 7.7). A binomial test confirmed that a significantly greater number of participants had a higher Bayes factor for the rational model than

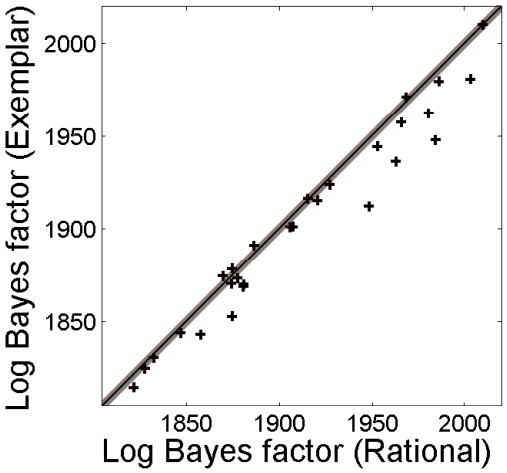

Figure 7.7: **Model fits**. Log Bayes factor for the rational model (relative to the veridical model) plotted against the log Bayes factor for the exemplar model. Each point represents a single participant from Experiment 1a, 1b or 2. Points below the diagonal favor the rational model. Error bars around the diagonal represent the 95 percent credible interval (see Gelman et al., 2004).

for the exemplar model ($p < 0.005$).

I then used the model fits to investigate the underlying representations posited by the two models. The exemplar model predicts that there should be 20 categories. In contrast, the fitted model prefers fewer categories (1.8 on average), demonstrating that the data are indeed better explained by assuming the simplicity principle.

7.5 General discussion

The experiments reported in this paper bring together two lines of research in cognitive psychology: the "simplicity principle" (a.k.a. "Occam's razor"; Chater and Vitányi, 2003) and the influence of categories on perception (Goldstone, 1995). I show that the simplicity bias towards merging perceptual categories together when their statistics are similar manifests itself in simple perceptual estimates. Participants tended to regularize estimates of trials of one color towards those of trials of another color if the stimulus distributions for the two colors had similar means (Experiments

1a and 1b) or overlapping tails (Experiment 2).

These findings are consistent with computational models that flexibly infer the number of categories from sensory inputs (Anderson, 1991; Gershman and Niv, 2010; Love et al., 2004; Sanborn et al., 2010). These models predict that new categories will only be postulated when stimulus statistics differ significantly; otherwise, the stimuli will be merged into a single category. This merging leads to regularization of perceptual estimates, such that perception of a new stimulus will be biased towards the mean of the merged distributions. I presented a rational adaptive categorization model that can predict the qualitative pattern of results and outperform alternative models in capturing subtle aspects of the behavioral data. Nonetheless, I have not yet fully mapped out the boundary conditions of the simplicity bias in my task, and so these data should be understood as initial explorations of my model's predictions rather than general statements about Occam's razor in categorical perception.

My results are consistent with other evidence that perception is influenced by unsupervised category learning. Gureckis and Goldstone (2008) asked participants to discriminate between pairs of stimuli that varied along two dimensions, and then in a second phase asked participants to classify these stimuli into two categories, where the classification boundary was determined by a single (attended) dimension. The stimuli were designed so that within each category, the stimuli fell into two sub-clusters on the basis of the second (unattended) dimension. Despite these sub-clusters being irrelevant for classification, participants were better able to discriminate between stimuli in the same category when they belonged to different sub-clusters. Thus, the underlying cluster structure of the stimuli systematically biased perception.

Although my study used numerical estimation as a paradigm for investigating perceptual biases, I was not interested in estimation *per se*: Only the relative estimation bias between conditions was relevant to my hypothesis. The speeded response requirement made it essentially impossible for participants to explicitly count the

number of circles on the screen, thus making past history (in particular, feedback from previous trials) a more influential factor in determining responses compared to the veridical number of circles. Nonetheless, my study may have implications for the study of number perception (Feigenson et al., 2004). In particular, my results suggest that numerical estimation is sensitive not only to the veridical numerosity, but can also be influenced by the distribution of numbers in recent experience. This points toward the existence of a more sophisticated number perception system that incorporates top-down knowledge about numerosity statistics.

I have interpreted my results in terms of Occam's razor, but alternative interpretations may also be possible. For example, an exemplar model (e.g., Kruschke, 1992; Nosofsky, 1986) that interpolates based on similarity between stimuli could also account for my results; however, I showed in Section 7.4.5, both quantitatively and qualitatively, that the rational model is a better explanation for the empirical data. Another viable alternative is a model in which the stimulus is assumed to have been drawn from one of two distributions (e.g., a mixture of Gaussians). In other words, the participant always assumes two distributions, but has uncertainty about which one generated the data. A potential problem with this account is that it assumes that participants already know the two distributions, whereas I am proposing that they infer them. Yet another possibility is that in the High mean alternative and High variance conditions, the difference between the two trial types was less salient; however, the fact that different colors were used for the different trial types argues against this interpretation.

A number of questions remain. For example, what are the sequential dynamics of category formation over the course of the experiment? Several previous studies have suggested that sequencing of exemplars plays an important role in unsupervised learning (Anderson, 1991; Clapper and Bower, 1994; Zeithamova and Maddox, 2009), and this factor may also come into play in my task. Although my experiments were not

designed to examine this factor directly, I reported significant sequential correlations in Experiment 1a, suggesting that the regularization effects I observed may operate over short timescales. Another question is whether the simplicity bias is itself subject to modulation by task factors. One possibility is that being repeatedly exposed to highly complex environments will lead to a greater tolerance for more complex category structures. Finally, it has been suggested that the hippocampus is a crucial neural substrate for category splitting and merging (Love and Gureckis, 2007; Gershman and Niv, 2010). Investigating this area's contributions to simple perceptual tasks like the one reported here is an important direction for future research.

7.5.1 Appendix: Particle filtering algorithm

The particle filter is an algorithm that approximates optimal Bayesian inference by updating an approximation to the posterior distribution over the assignment of trials to categories as each observation arrives. This sequential online nature makes it suitable for modeling the dynamics of human learning in my experiments. Similar process models have previously been applied to animal (Daw and Courville, 2008; Gershman and Niv, 2010) and human (Brown and Steyvers, 2009; Frank et al., 2010; Sanborn et al., 2010) learning, although the generative assumptions of those models differ from my own.[2]

The particle filtering algorithm maintains a set of L samples $z_{t-1}^{(1:L)}$ distributed approximately according to the posterior, $P(z_{t-1}|\mathbf{c}_{1:t-1}, \mathbf{x}_{1:t-1})$. These samples are updated after observing x_t and c_t by drawing $z_t^{(l)}$ for $l = 1, \ldots, L$ from $P(z_t^{(l)} = k) =$

[2]While the particle filter provides a plausible mechanism by which participants might perform approximate Bayesian inference, it is by no means the only one. I present it merely as an example of how the approximation might be accomplished, without committing to any particular process-level account.

$\frac{w_k}{\sum_k w_k}$, where

$$w_k = \sum_{l=1}^{L} P(c_t|z_t = k, \mathbf{z}_{1:t-1}^{(l)}, \mathbf{c}_{1:t-1}) P(x_t|z_t = k, \mathbf{z}_{1:t-1}^{(l)}, \mathbf{x}_{1:t-1}) P(z_t = k|\mathbf{z}_{1:t-1}^{(l)}). \quad (7.12)$$

Drawing samples in this way produces a Monte Carlo approximation to the posterior (Doucet et al., 2001). As $L \to \infty$, this approximation will converge to the true posterior (see Chapter 2).

The particle filter can also be used to estimate the number of circles x_t given noisy sensory input y_t (before feedback):

$$\begin{aligned} \mathbb{E}[x_t|y_t, \mathbf{x}_{1:t-1}, \mathbf{c}_{1:t}] &= \sum_{\mathbf{z}_{1:t}} \int_x x P(x_t = x, \mathbf{z}_{1:t}|y_t, \mathbf{x}_{1:t-1}, \mathbf{c}_{1:t}) dx \\ &\approx \frac{1}{L} \sum_{l=1}^{L} \frac{\sum_k q_k^{(l)} m_k^{(l)}}{\sum_k q_k^{(l)}}, \end{aligned} \quad (7.13)$$

where

$$q_k^{(l)} = P(c_t|z_t = k, \mathbf{z}_{1:t-1}^{(l)}, \mathbf{c}_{1:t-1}) P(z_t = k|\mathbf{z}_{1:t-1}^{(l)}) \quad (7.14)$$

is the posterior weight assigned to category k and

$$\begin{aligned} m_k^{(l)} &= \int_x x P(x_t = x|y_t, z_t = k, \mathbf{z}_{1:t-1}^{(l)}, \mathbf{x}_{1:t-1}) dx \\ &= \int_x x \frac{P(y_t|x_t = x) P(x_t = x|z_t = k, \mathbf{z}_{1:t-1}^{(l)}, \mathbf{x}_{1:t-1})}{P(y_t, z_t = k, \mathbf{z}_{1:t-1}^{(l)}, \mathbf{x}_{1:t-1})} dx \end{aligned} \quad (7.15)$$

is the prediction of x_t for category k. I know of no closed-form expression for $m_k^{(l)}$, but I can obtain a very accurate numerical approximation. In my implementation, I set $L = 100$, but the results are not sensitive to this choice.

Chapter 8

Neural context reinstatement and memory misattribution

In Chapter 4, I introduced an important principle of memory: retrieval renders memories susceptible to disruption (or enhancement) by amnestic agents (Lee et al., 2006; Misanin et al., 1968; Nader et al., 2000) or new learning (Chan et al., 2009; Galluccio and Rovee-Collier, 2005; Spear, 1973). I proposed a computational model based on a combination of storage and retrieval mechanisms to explain some of the most prominent results from this literature. However, recent results in humans and rats highlight aspects of reactivation-based memory disruption that are challenging for my model. In this chapter, I review these findings and discuss an alternative, retrieval-based theoretical perspective my colleagues and I have proposed (Sederberg et al., 2011). I then present neural evidence (in humans) supporting the retrieval-based theory. In Section 8.4, I return to the latent cause perspective and suggest how these perspectives might be connected.

8.1 Asymmetric memory misattributions in the Hupbach paradigm

Hupbach et al. (2007) introduced a reactivation-based behavioral paradigm to study episodic memory misattributions in humans. The experiment consisted of three sessions, each separated by 48 hours. On the first day, subjects studied a set of objects that were pulled (one by one) out of a blue basket (list 1, L1). In session 2, subjects returned to the lab and studied a new set of objects (L2), this time spread out over a table. The key manipulation was that, prior to studying L2, some participants were given a reminder about session 1. Specifically, in the reminder condition, the same experimenter who was present during session 1 took the participants back to the same room that was used during session 1; the experimenter showed the blue basket to the subject and asked whether they remembered studying the items in the basket (subjects were stopped if they started to recall any specific items out loud). After the reminder, subjects studied the L2 items (in the same room as session 1). In the no-reminder condition, a new experimenter took the subjects to a new room to study the L2 items. Finally, in session 3 subjects were asked to free recall items from either L1 or L2.

The main finding of Hupbach et al. (2007) was an asymmetric pattern of intrusions in the reminder condition: when asked to recall items from L1, subjects tended to erroneously recall items from L2, but when asked to recall items from L2, subjects rarely erroneously recalled items from L1. Subjects in the no-reminder condition did not show an asymmetric pattern of misattributions, and showed a low level of misattributions across both lists. Hupbach et al. (2007) also found that this effect only occurred when the test session was given 48 hours after session 2; subjects tested immediately after studying L2 did not intrude a substantial number of L2

items when recalling L1.[1] Follow-up studies have provided additional extensions to and constraints on this effect: the effect can be replicated in children (Hupbach et al., 2011) and rats (Jones et al., 2012); the effect occurs with a source memory test (Hupbach et al., 2009); spatial context is a necessary and sufficient reminder (Hupbach et al., 2008), but only when it is unfamiliar (Hupbach et al., 2011).

8.2 A theoretical perspective: the Temporal Context Model

These findings are somewhat puzzling from the perspective of the latent cause framework. If the reminder in Hupbach's paradigm operates in a similar manner to the reminders in reconsolidation experiments, then the models presented in Chapters 3 and 4 would (mutatis mutandis) predict that L2 items will be assigned to the same latent cause as L1 items—that is, only a single memory trace will be formed. But such an explanation seems at odds with the finding of asymmetric intrusions; a single memory trace account would predict symmetric intrusions.

I now briefly describe an alternative model, designed precisely to deal with human free recall experiments, known as the *Temporal Context model* (TCM; Howard and Kahana, 2002; Sederberg et al., 2008). This model is specified at the algorithmic level rather than at the computational level (but see Gershman et al., 2012). It embodies a number of mechanistic principles that are important for modeling human memory, but which are not currently built into my rational analysis of memory, such as the notion of a drifting "temporal context" (see below).

Figure 8.1 shows the model as a two-layer neural network. In TCM, each item is represented by a unit vector. We refer to this representation as the "item layer." As items are presented, these representations feed into a "context" layer via a linear map-

[1] Though evidently this finding does not extend to rats; see Jones et al. (2012).

ping. The context activity represents a superposition of recent item representations (i.e., a "temporal context"). The context-to-item weights, updated using Hebbian learning, allow the model to retrieve item representations given a context pattern; similarly, the item-to-context weights allow the model to retrieve a context pattern given an item pattern. When an item is recalled, its study context is reinstated on the context layer; this serves as a retrieval cue for other items with similar temporal contexts. In Sederberg et al. (2011), we showed that this model could account for most of the phenomena discovered by Hupbach and her colleagues.

TCM is able to capture asymmetric intrusions in Hupbach's paradigm as a consequence of item-context binding. In the reminder condition, the L1 context unit is associated with both L1 and L2 items, whereas the L2 context unit is always bound only to the L2 items. Consequently, cuing with the L1 context unit triggers recall of both L1 and L2 items, whereas cuing with the L2 context unit primarily triggers recall of L2 items. Sederberg et al. (2011) simulated several other aspects of the Hupbach paradigm, which I omit here (along with all the technical details of the model and simulations).

A more direct test of the TCM account would be if we could directly measure L1 context reinstatement during L2 study and ask whether the degree of reinstatement predicts which L2 items will be subsequently misattributed to L1. Of course, we cannot do this with behavior, but (as I describe in the next section) functional brain imaging provides a window onto mental processes which we can exploit to investigate this question.

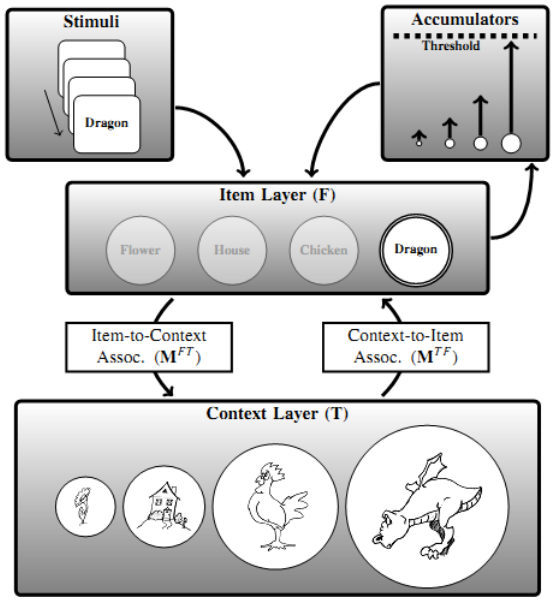

Figure 8.1: **Schematic of the Temporal Context Model**. The retrieval mechanism in this particular instantiation is a set of leaky competing accumulators. Reprinted from Sederberg et al. (2008).

8.3 Experiment: functional brain imaging of the Hupbach paradigm

The logic of this experiment was to create a situation in which context reinstatement could be effectively measured. To do this, I leveraged the existence of a cortical area that responds selectively to scene, known as the *parahippocampal place area* (PPA; Epstein and Kanwisher, 1998). By interposing scenes during the intertrial interval (ITI) separating L1 items, I assumed that temporal context during L1 study would have a preponderance of scene-related activity. Using data from a scene localizer run (described below), I trained a multivariate pattern classifier to detect scene-related activity. During L2 study, I then used the classifier to measure scene-related activity, and used this as a proxy for context reinstatement. I could then sort the L2 items according to whether they were subsequently misattributed to L1 in a later recognition memory test, analogous to "subsequent memory" analyses widely used

in neuroimaging of memory (Paller and Wagner, 2002). The hypothesis, based on the TCM account of Hupbach's findings described above, was that scene context reinstatement would predict which items would subsequently be misattributed to L1.

8.3.1 Materials and Methods

Subjects. Fourteen right-handed subjects participated in the study. All were free of neurological or psychiatric disease, and fully consented to participate.

Stimuli and task. The experiment consisted of three sessions, each separated by 48 hours. All sessions took place inside the fMRI scanner. Stimuli were presented using the Psychophysics Toolbox (Brainard, 1997). The experimental design is shown schematically in Figure 8.2. During the first session, subjects studied a list of 20 items (object pictures), presented sequentially on the computer via a projection system that reflected the images onto a mirror in the bore of the magnet. Each item was presented for 2 s (highlighted by a green frame), followed by an ITI randomly jittered between 4 and 7 s. During session 1, the ITI was filled by a continuous sequence of random scene images (duration: 1 s); during session 2, the ITI was a blank screen. The list was presented 4 times in random order, each time followed by a free recall task in which subjects were asked to verbally recall the names of objects studied in the list. The free recall task was performed inside the scanner, between functional scans.

Prior to the beginning of session 2, I gave subjects a "reminder" of session 1, analogous to the reminders used in previous studies (Hupbach et al., 2007, 2008, 2009, 2011). Specifically, I asked subjects to recall the general procedure during session 1. Invariably, they described studying and recalling a list of items repeatedly; they did not describe any of the specific study items, suggesting that the reminder predominantly reactivated "context" memories rather than memories of specific items. Note that I did not include any subjects in a no-reminder condition. The rest of session 2 proceeded in an identical manner to session 1, with subjects studying a new list of

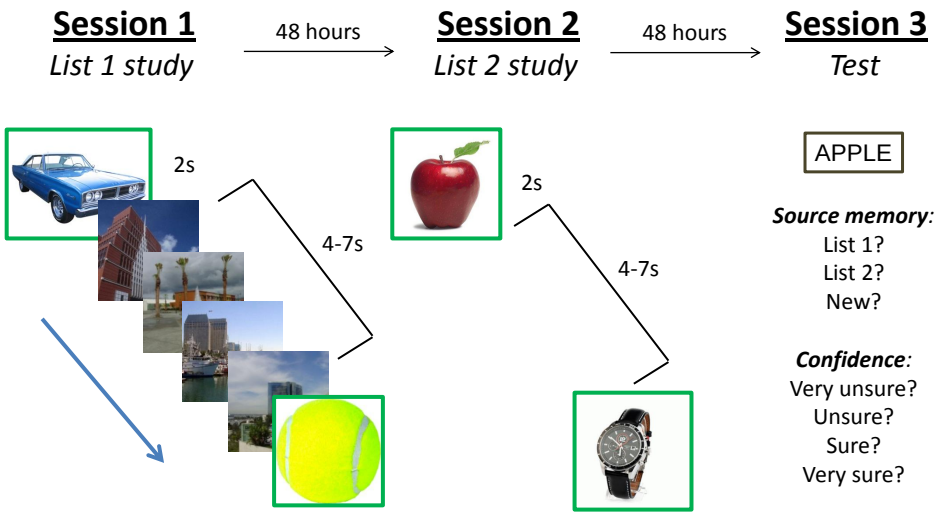

Figure 8.2: **Experimental design**.

20 items 4 times, with free recall after each list repetition.

During session 3, subjects performed a recognition task in which they were asked to judge whether an item (presented as an object name) was studied in session 1, session 2 or neither (i.e., a new item). After each recognition judgment, subjects were asked to rate their confidence on a 4-point scale (sure old, not sure old, not sure new, sure new). Responses were recorded using a button box.

Following the recognition task, I ran a scene "localizer" run, in which subjects viewed alternating mini-blocks of scene and phase-scrambled scene images. Each mini-block consisted of 8 images, each presented for 500 ms and separated by an ITI of 1.5 s. A total of 16 mini-blocks were presented, each separated by 12 seconds. To keep subjects focused, they were asked to press a button each time they detected a repeated image.

fMRI data acquisition. Data were acquired using a 3T Siemens Allegra scanner with a volume head coil. I collected four functional runs in sessions 1 and 2 and two functional runs in session 3 with a T2*-weighted gradient-echo EPI sequence (35 oblique axial slices, 3.5 × 3.5 mm inplane, 3.5 mm thickness; TE=28 ms; TR=2000 ms; FA=71°; matrix=64 × 64; field of view=224 mm). I collected two anatomical

runs for registration across sessions and across subjects to standard space: a coplanar T1-weighted FLASH sequence (35 oblique axial slices, $3.5 \times 3.5mm$ in plane, 3.5 mm thickness; TE=4.60 ms; TR=400 ms; FA=90°; matrix=64 × 64; field of view=224 mm) and a high-resolution 3D T1-weighted MPRAGE sequence (176 1 mm sagittal slices; TE=3.34 ms; TR=2500 ms; FA=7°; matrix=256×256; field of view=256 mm). A FLASH image was acquired for each session, while only a single MPRAGE was acquired per subject.

fMRI data preprocessing. Preprocessing was performed using Statistical Parametric Mapping software (SPM8; Wellcome Department of Imaging Neuroscience, Institute of Neurology, London, UK). Images were realigned to correct for subject motion and coregistered across sessions using an affine transformation of the FLASH images. The data were then high-pass filtered with a cutoff period of 128 s. No spatial normalization or smoothing were applied to the data.

Region of interest (ROI) selection. A general linear model (GLM) was fit to the localizer data for each subject with "scene" and "scrambled scene" regressors convolved with the canonical hemodynamic response function. The six scan-to-scan motion parameters produced during motion correction were also included as additional nuisance regressors in the GLM to account for residual effects of subject movement. A t-statistic map was then created for the scene > scrambled scene contrast, thresholded at $p < 0.001$ (uncorrected). Bilateral clusters corresponding anatomically to the PPA in the posterior parahippocampal/collateral sulcus region (Epstein and Kanwisher, 1998; Epstein et al., 1999) were selected as functional regions of interests for each subject individually.

Multivariate pattern analysis. PPA activity for each mini-block within the localizer run was averaged into a single image. I also included a 3rd "rest" class obtained by averaging the activation during the second half of the inter-block interval. These images were entered into an l_2-regularized multinomial logistic regression classifier,

trained to predict scene vs. scrambled scene vs. rest labels. The regularization parameter was set to 0.1, but we found that the results were insensitive to varying this parameter over 3 orders of magnitude. The trained classifier was then used to predict scene activity (quantified as the probability of assigning an image to the scene category) during the study and test runs. The rationale for including the rest class was that it can capture activity associated with "mind-wandering" activity that is not due to changes in category-specific activation. I obtained very similar results if we excluded the rest class.

8.3.2 Results

Behavior. As shown in Figure 8.3, I replicated the asymmetric pattern of intrusions found by Hupbach and colleagues: L2 items were more frequently misattributed to L1 items, than vice versa $[t(13) = 4.00, p < 0.002]$. The overall level of false alarms, where novel items were judged as old, was low (median = 2). My replication of Hupbach's results is notable, in that this is the first time that the paradigm has been adapted to a standard list learning setup (not to mention in a scanner). Thus, I can comfortably dismiss concerns that the effect was idiosyncratic to the somewhat unusual experimental conditions used by Hupbach in her studies.

Imaging results. I evaluated the classifier's predictions for scene activity at several time points before and after the trial onset of L2 items in session 2, as shown in Figure 8.4A. These predictions were sorted according to whether L2 items were subsequently correctly attributed to L2 (red line) or misattributed to L1 (blue line). The results show that at $t = -2$, scene activity was significantly higher for L2 items subsequently misattributed to L1 compared to those subsequently correctly attributed to L2 $[t(13) = 2.71, p < 0.02]$. The fact that this effect occurs before the trial onset is important: in order for items to be bound to the scene context, this context must,

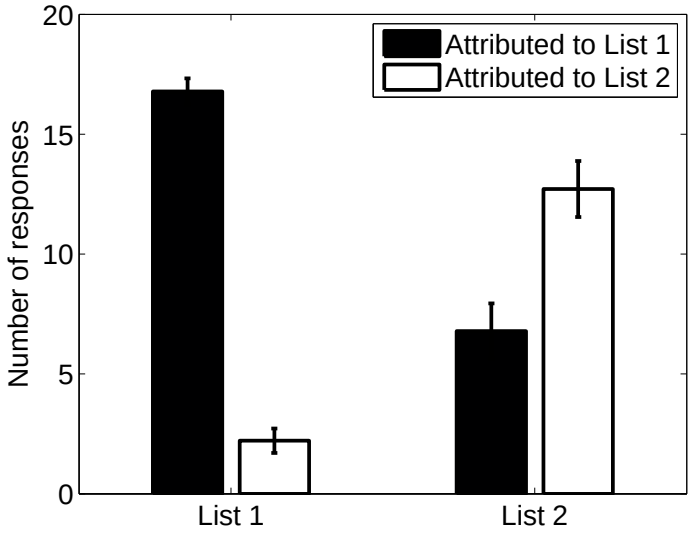

Figure 8.3: **Behavioral results**.

according to TCM, be reinstated before the item presentation.[2] Thus, the classifier predictions provide compelling support for TCM's interpretation of the asymmetric intrusions.

Figure 8.4B shows the same analysis for scene activity at test, this time including L1 items subsequently attributed to L1 (very few L1 items were misattributed to L2). In this case, I found no significant differences between any of the conditions. This may be due to the fact that subjects had to make two responses immediately after being presented with the test cue, which might interfere with the classifier's ability to decode scene activation.[3]

To obtain a more fine-grained picture of how scene activation related to memory performance, I next examined whether the scene activation was predictive of parametric differences in response confidence. I first converted the confidence ratings to an "unfolded" scale, where -4 represents "sure L2" and $+4$ represents "sure L1." I then fit, for each time point during the trial, a linear regression model with the scene

[2]Because I had an *a priori* hypothesis (based on TCM) about the time point at which the effect should be observed, I did not correct for multiple comparisons.

[3]When the rest class was excluded, I found that correctly attributed L1 items show greater scene activity than correctly attributed L2 items at $t = +2$ [$t(13) = 2.30, p < 0.05$].

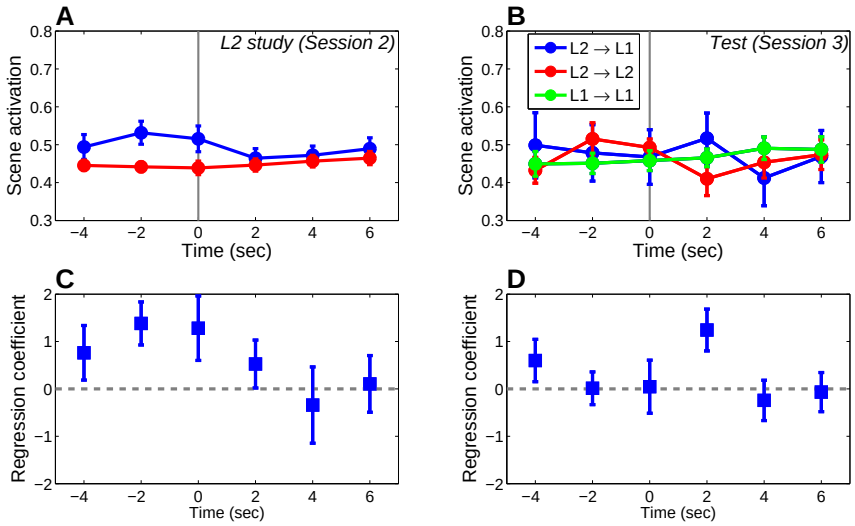

Figure 8.4: **Imaging results**. (*A*) Time course of scene activation during study of L2 in session 2. The blue line represents L2 items subsequently misattributed to L1, and the red line represents L2 items correctly attributed to L2. The vertical line represents the trial onset. Scene activation was measured by the logistic regression classifier's prediction of the probability that the mental state of "scene" is present. (*B*) Time course of scene activation during the recognition test (session 3). The green line represents L1 items correctly attributed to L1. (*C*) Linear regression between scene activation during each time point of L2 study trials and "unfolded" confidence at test for misattributed L2 items. (*D*) Linear regression between scene activation during each time point of L2 test trials and "unfolded" confidence at test for misattributed L2 items. Error-bars represent standard error of the mean.

activation as the predictor and unfolded confidence as the response variable.

Figure 8.4C shows the results for L2 study: unfolded confidence increases monotonically as a function of scene activation for prestimulus time points, as indicated by positive regression coefficients for these timepoints. Thus, the stronger the reinstatement of L1 context, the more confident subjects were that L2 items were seen in L1. To assess the significance of these results in a within-subject manner, I performed a t-test on the regression coefficients at each time point against zero. This analysis revealed a significant within-subject effect at $t = -2$ $[t(13) = 3.04, p < 0.01]$. This finding fits with TCM's prediction that items are bound to the context reinstated just *before* the item presentation.

At test (Figure 8.4D), we found that scene reactivation was significantly elevated for subsequently misattributed $L2$ items at $t = +2$ $[t(13) = 2.82, p < 0.05]$. Thus, the degree of context reinstatement following item retrieval at test also predicts the confidence with which subjects judge that an L2 item was studied in L1. This is consistent with TCM's prediction that the study context will be reinstated following a retrieval cue.

8.4 Discussion

Hupbach et al. (2007) reported one of the most intriguing findings from the reconsolidation literature—an asymmetric pattern of memory misattributions following a reminder treatment. Sederberg et al. (2011) proposed a theoretical explanation for this finding based on TCM, which has successfully modeled many list learning phenomena (Howard and Kahana, 2002; Sederberg et al., 2008). TCM makes a very clear prediction: misattributions should occur when L1 context is reinstated prior to L2 study. More precisely, L2 items will be misattributed if they are bound to L1 context. I tested and confirmed this prediction using fMRI: context, measured neurally by the reactivation of scene-related cortical patterns, predicted which L2 items would be subsequently misattributed to L1. Moreover, these predictions could also parametrically predict the degree of confidence in these misattributions.

These findings fit well with the recent literature investigating the role of neural context reinstatement in memory tasks. For example, Polyn et al. (2005) showed that category-specific neural activation precedes the recall of items and can predict the category of item recalled. Johnson et al. (2009) used this method to show that context reinstatement occurs during recollection-based judgments, and Kuhl et al. (2011) showed that the degree of context reinstatement predicted the outcome of memory competition. The results reported in this chapter are unique in demonstrating a link

between context reinstatement and memory misattributions following a reminder.

While I have focused on TCM in this chapter, the question remains how these results might be reconciled with the latent cause framework, which (as discussed above) does not predict asymmetric misattributions. Part of the problem is that the latent cause theories described in previous chapters assume that in certain cases memories will overwrite each other, whereas TCM stores an indelible trace of all items and contexts, relying on retrieval interference to produce misattributions. One possibility is that in a fully Bayesian framework there is no overwriting—that is, there is always some probability that items were generated by different latent causes. I do not know yet if such a model could predict asymmetric misattributions, but it seems unlikely.

TCM's explanation in terms of context reinstatement is simple and persuasive, so perhaps the answer is in looking for analogies between the latent cause framework and temporal context. In previous work, I have provided a normative interpretation of TCM (Gershman et al., 2012), but this interpretation does not involve latent causes. A different way to look at the connection between latent causes and temporal context is to view the context representation as a kind of distributed belief about latent causes. Loosely speaking, TCM implements a "soft clustering" of items into different temporal contexts, analogous to the organizational role of latent causes in partitioning observations. Thus, the probability that an item should be assigned to a particular cluster may correspond to the degree of context activation. My colleagues and I explored a version of this idea in an earlier paper (Socher et al., 2009). The proper treatment (if one exists) of the connection between context-based theories like TCM and my latent cause framework remains a stimulating unsolved problem.

Chapter 9

Conclusion

William James once remarked that "the art of remembering is the art of thinking" (James, 1901). If we take this statement seriously, we must ask: what is the nature of this thinking? In this thesis, I have advanced the viewpoint that the thinking underlying memory is essentially statistical in nature, and the memories formed through this thought process reflect inductive inferences about the latent variables in the environment. In particular, I introduced the idea that memory traces correspond to inferences about latent causes; these inferences serve to organize observations into "clusters" and guide predictions about future observations.

9.1 Lessons learned

Below I summarize the main lessons learned from the work described in this thesis.

Lesson 1: memory traces correspond to inferences about latent causes. In Chapter 3, I elaborated a detailed theory of how memory traces can be formalized rationally in terms of inferences about latent causes. Intuitively, if you observe A and infer that it was generated by the same latent cause as B, then your memory for B should be modified by observation A. This basic idea, along with its elaborations, is able to account for a large number of empirical findings, as demonstrated throughout

this thesis.

Lesson 2: the modification rules for memories obey a rich set of inductive biases. I have described several "inductive biases" which enter into the Bayesian latent cause framework as "priors." These inductive biases aid learning by constraining the possible interpretations of observational data, allowing strong inferences to be made from very small amounts of data (Griffiths et al., 2010). I hypothesized (and presented experimental evidence for) several inductive biases: causes tend to be correlated in time (Chapter 4), change tends to be gradual (Chapter 5), and a small number of latent causes are more likely *a priori* (Chapter 7).

Lesson 3: the dynamical interplay between associative and structure learning determines when memories are modified. My model of reconsolidation (Chapter 4) hinged crucially on the dynamics of associative learning (updating of CS-US associations) and structure learning (inference about latent causes). The basic lesson from this is that as associative learning decrements weights following the first extinction trial, the statistical pressure for the structure learning system to infer a new latent cause decreases. In other words, there are two ways to respond to an extinction-induced prediction error (decreasing the CS-US association or inferring a new latent cause), and the timing of trials exerts exquisite control over which of these responses prevails.

Lesson 4: gradual change facilitates memory modification. As I showed experimentally in Chapters 5 and 6, memory modification can be promoted by gradually changing the observational statistics, which increases the posterior probability that observations at different timepoints were generated by the same latent cause.

Lesson 5: context reinstatement predicts memory misattribution. Using fMRI, I presented evidence for the idea that mental context reinstatement determines the degree to which new observations modify old memories (Chapter 8). This contextual reinstatement can be interpreted as a signal that an earlier latent cause is now

active again. Latent causes represent one statistical interpretation of "context."

9.2 One model to rule them all?

This thesis describes many models. Why can't there be just one model to rule them all? There are several different ways to answer this question. One is that the different phenomena modeled in this thesis entail different computational demands, although they share a common theoretical core. The demands of a Pavlovian fear conditioning task (Chapters 3-5) are intrinsically different from those of visual memory (Chapter 6) or perceptual estimation (Chapter 7). Thus, the models I presented used different parameterizations to capture unique demands of various tasks.

Another answer is that the latent cause model is not sufficiently developed for certain tasks, such as the human memory task described in Chapter 8 (the Hupbach reminder paradigm). In that chapter, I used the well-established Temporal Context Model (Howard and Kahana, 2002; Sederberg et al., 2008) as an organizing framework for thinking about the Hupbach paradigm. TCM was specifically designed to model human list learning (and in particular free recall tasks). There is now extremely compelling behavioral and neural evidence for its explanatory power (Polyn and Kahana, 2008). Moreover, my colleagues and I have already published a theoretical paper showing how TCM can account for many of Hupbach's findings (Sederberg et al., 2011). For these reasons, it seemed natural to view my new data through the lens of TCM, and in fact my experimental design and predictions were specifically motivated by TCM. In Chapter 8 I ventured some tentative speculations about the relationship between TCM and the latent cause framework, but this is a situation where the latent cause framework requires further development.

A major lacuna of the Pavlovian conditioning model described in Chapter 4 is that I was unable to comprehensively capture the results of my gradual conditioning

experiments (Chapter 5). The basic problem is that the model does not adequately distinguish the gradual and gradual reverse conditions. Indeed, these conditions were remarkably similar in terms of the CS-US schedule. I have tried many different parameter settings, as well as a particle filter implementation, but none of the simulations provided a satisfying fit.

My intuition is that all the models described in my thesis fail to capture a potentially important aspect of the experimental paradigm: change is *directional*. When we watch an apple fall from a tree, we know that the change process is essentially unidirectional, not Brownian motion as the model of Chapter 6 assumes. I think that something similar is happening in gradual extinction: the rat makes an inference that the CS-US contingency is decreasing monotonically. By contrast, the rat infers in the gradual reverse condition that the CS-US contingency is *increasing* monotonically. These monotonicity inferences lead to powerful predictions about future trials, much more powerful than the predictions licensed by Brownian motion. The weakness of the latter model's predictions is a major reason for the failure of this model to differentiate between the gradual and gradual reverse conditions. Thus, a model that incorporates inference over monotonic change dynamics could potentially explain the dramatic differences between conditions shown in Chapter 4.

9.3 Envoi

The computational framework described in this thesis is still in its infancy. Perhaps the most urgent and exciting direction is to develop clinical applications of some of these ideas. For example, the results of gradual extinction suggest that post-traumatic stress disorder or addiction could be treated by gradually reducing the association between the traumatic (or addictive) event and the stimuli with which it co-occurred. The greatest contribution a computational theory can make to clinical

treatment is the ability to *quantitatively* predict outcomes. In the near future, I see the development of quantitatively precise theories of learning and memory as playing an increasingly important role in treating mental disorders.

Bibliography

Aimone, J., Wiles, J., and Gage, F. (2006). Potential role for adult neurogenesis in the encoding of time in new memories. *Nature neuroscience*, 9(6):723–727.

Alberini, C. (2007). Reconsolidation: The Samsara of memory consolidation. *Debates in Neuroscience*, 1(1):17–24.

Aldous, D. (1985). Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII*, pages 1–198. Springer, Berlin.

Amaral, O., Osan, R., Roesler, R., and Tort, A. (2008). A synaptic reinforcement-based model for transient amnesia following disruptions of memory consolidation and reconsolidation. *Hippocampus*, 18(6):584–601.

Anderson, J. (1990). *The adaptive character of thought*. Lawrence Erlbaum.

Anderson, J. (1991). The adaptive nature of human categorization. *Psychological Review*, 98(3):409–429.

Antoniak, C. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2:1152–1174.

Barnes, J. and Underwood, B. (1959). "fate" of first-list associations in transfer theory. *Journal of Experimental Psychology*, 58(2):97–105.

Bayer, H. and Glimcher, P. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, 47(1):129–141.

Becker, S. (2005). A computational principle for hippocampal learning and neurogenesis. *Hippocampus*, 15(6):722–738.

Biedenkapp, J. and Rudy, J. (2004). Context memories and reactivation: constraints on the reconsolidation hypothesis. *Behavioral neuroscience*, 118(5):956.

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Springer.

Bissière, S., Humeau, Y., and Luthi, A. (2003). Dopamine gates ltp induction in lateral amygdala by suppressing feedforward inhibition. *Nature Neuroscience*, 6(6):587–592.

Blaisdell, A. P., Sawa, K., Leising, K. J., and Waldmann, M. R. (2006). Causal reasoning in rats. *Science*, 311:1020–1022.

Blei, D. and Frazier, P. (2011). Distance dependent chinese restaurant processes. *Journal of Machine Learning Research*, 12:2461–2488.

Blumenfeld, B., Preminger, S., Sagi, D., and Tsodyks, M. (2006). Dynamics of memory representations in networks with novelty-facilitated synaptic plasticity. *Neuron*, 52(2):383–394.

Boehner, P., editor (1957). *Ockham: Philosophical Writings*. Hackett Publishing, Nelson, Edinburgh.

Bouton, M. and Bolles, R. (1979a). Contextual control of the extinction of conditioned fear. *Learning and Motivation*, 10(4):445–466.

Bouton, M. and Bolles, R. (1979b). Role of conditioned contextual stimuli in reinstatement of extinguished fear. *Journal of Experimental Psychology*, 5:368–378.

Bouton, M. and King, D. (1983). Contextual control of the extinction of conditioned fear: Tests for the associative value of the context. *J Exp Psychol*, 9:248–265.

Bouton, M., Woods, A., and Pineño, O. (2004). Occasional reinforced trials during extinction can slow the rate of rapid reacquisition. *Learning and Motivation*, 35(4):371–390.

Bouton, M. E. (1993). Context, time, and memory retrieval in the interference paradigms of Pavlovian learning. *Psychological Bulletin*, 114:80–99.

Bouton, M. E. (2004). Context and behavioral processes in extinction. *Learning and Memory*, 11:485–494.

Bradley, P. and Galal, K. (1988). State-dependent recall can be induced by protein synthesis inhibition: Behavioural and morphological observations. *Developmental Brain Research*, 40(2):243–251.

Brainard, D. (1997). The psychophysics toolbox. *Spatial vision*, 10(4):433–436.

Brandon, S. E., Vogel, E. H., and Wagner, A. R. (2000). A componential view of configural cues in generalization and discrimination in Pavlovian conditioning. *Behavioral Brain Research*, 110:67–72.

Briggs, J. and Riccio, D. (2007). Retrograde amnesia for extinction: Similarities with amnesia for original acquisition memories. *Learning & Behavior*, 35(3):131.

Brogden, W. (1939). Sensory pre-conditioning. *Journal of Experimental Psychology*, 25:323–332.

Brown, A. (1976). Spontaneous recovery in human learning. *Psychological Bulletin*, 83(2):321–338.

Brown, A. (2002). Consolidation theory and retrograde amnesia in humans. *Psychonomic Bulletin & Review*, 9(3):403.

Brown, G., Neath, I., and Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, 114(3):539–576.

Brown, S. and Steyvers, M. (2009). Detecting and predicting changes. *Cognitive Psychology*, 58(1):49–67.

Capaldi, E. (1957). The effect of different amounts of alternating partial reinforcement on resistance to extinction. *The American Journal of Psychology*, pages 451–452.

Chan, J., Thomas, A., and Bulevich, J. (2009). Recalling a witnessed event increases eyewitness suggestibility: the reversed testing effect. *Psychological Science*, 20(1):66–73.

Chan, W., Leung, H., Westbrook, R., and McNally, G. (2010). Effects of recent exposure to a conditioned stimulus on extinction of pavlovian fear conditioning. *Learning & Memory*, 17(10):512–521.

Chater, N. and Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in cognitive sciences*, 7(1):19–22.

Clapper, J. and Bower, G. (1994). Category invention in unsupervised learning. *Journal of experimental psychology: learning, memory and cognition*, 20:443–443.

Clem, R. and Huganir, R. (2010). Calcium-permeable ampa receptor dynamics mediate fear memory erasure. *Science*, 330(6007):1108–1112.

Colgin, L., Moser, E., and Moser, M. (2008). Understanding memory through hippocampal remapping. *Trends in neurosciences*, 31(9):469–477.

Collins, A. and Koechlin, E. (2012). Reasoning, learning, and creativity: Frontal lobe function and human decision-making. *PLoS Biology*, 10(3):e1001293.

Corbit, L. H. and Balleine, B. W. (2000). The role of the hippocampus in instrumental conditioning. *Journal of Neuroscience*, 20:4233–4239.

Costanzi, M., Cannas, S., Saraulli, D., Rossi-Arnaud, C., and Cestari, V. (2011). Extinction after retrieval: Effects on the associative and nonassociative components of remote contextual fear memory. *Learning & Memory*, 18(8):508–518.

Courville, A. (2006). *A latent cause theory of classical conditioning*. PhD thesis, Pittsburgh, PA, USA.

Courville, A., Daw, N., Gordon, G., and Touretzky, D. (2004). Model uncertainty in classical conditioning. In Thrun, S., Saul, L., and Schölkopf, B., editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA.

Courville, A., Daw, N., and Touretzky, D. (2002). Similarity and discrimination in classical conditioning: A latent variable account. In Saul, L. K., Weiss, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 17*, pages 313–320, Cambridge, MA. MIT Press.

Courville, A., Daw, N., and Touretzky, D. (2006). Bayesian theories of conditioning in a changing world. *Trends in cognitive sciences*, 10(7):294–300.

Crowder, R. (1976). *Principles of Learning and Memory*. Lawrence Erlbaum.

Daw, N. and Courville, A. (2008). The pigeon as particle filter. In Platt, J., Koller, D., Singer, Y., and Roweis, S., editors, *Advances in Neural Information Processing Systems 20*, pages 369–376. MIT Press, Cambridge, MA.

Daw, N., Courville, A., and Touretzky, D. (2006). Representation and timing in theories of the dopamine system. *Neural Computation*, 18(7):1637–1677.

Dayan, P. and Abbott, L. (2001). *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. MIT Press.

Dayan, P., Kakade, S., and Montague, P. R. (2000). Learning and selective attention. *Nature Neuroscience*, 3 Suppl:1218–1223.

Dayan, P. and Long, T. (1998). Statistical models of conditioning. In *Advances in Neural Information Processing Systems 10*, pages 117–123, Cambridge, MA, USA. MIT Press.

Dayan, P., Niv, Y., Seymour, B., and Daw, N. (2006). The misbehavior of value and the discipline of the will. *Neural Networks*, 19(8):1153–1160.

Debiec, J., LeDoux, J., and Nader, K. (2002). Cellular and systems reconsolidation in the hippocampus. *Neuron*, 36(3):527–538.

Delamater, A. (2004). Experimental extinction in pavlovian conditioning: Behavioural and neuroscience perspectives. *Quarterly Journal of Experimental Psychology Section B*, 57(2):97–132.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Denève, S., Duhamel, J., and Pouget, A. (2007). Optimal sensorimotor integration in recurrent cortical networks: a neural implementation of kalman filters. *The Journal of neuroscience*, 27(21):5744–5756.

Devietti, T., Conger, G., and Kirkpatrick, B. (1977). Comparison of the enhancement gradients of retention obtained with stimulation of the mesencephalic reticular formation after training or memory reactivation. *Physiology & Behavior*, 19(4):549–554.

Doucet, A., De Freitas, N., and Gordon, N. (2001). *Sequential Monte Carlo Methods in Practice*. Springer Verlag.

Doyère, V., Debiec, J., Monfils, M., Schafe, G., and LeDoux, J. (2007). Synapse-

specific reconsolidation of distinct fear memories in the lateral amygdala. *Nature Neuroscience*, 10(4):414–416.

Duvarci, S. and Nader, K. (2004). Characterization of fear memory reconsolidation. *Journal of Neuroscience*, 24(42):9269.

Eisenberg, M., Kobilo, T., Berman, D., and Dudai, Y. (2003). Stability of retrieved memory: inverse correlation with trace dominance. *Science*, 301(5636):1102.

Epstein, R., Harris, A., Stanley, D., and Kanwisher, N. (1999). The parahippocampal place area:: Recognition, navigation, or encoding? *Neuron*, 23(1):115–125.

Epstein, R. and Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, 392(6676):598–601.

Eysenck, H. (1968). A theory of the incubation of anxiety/fear responses. *Behaviour Research and Therapy*, 6(3):309–321.

Fearnhead, P. (2004). Particle filters for mixture models with an unknown number of components. *Journal of Statistics and Computing*, 14:11–21.

Feigenson, L., Dehaene, S., and Spelke, E. (2004). Core systems of number. *Trends in cognitive sciences*, 8(7):307–314.

Feldman, J. (2003). The simplicity principle in human concept learning. *Current Directions in Psychological Science*, pages 227–232.

Foster, D. J., Morris, R. G., and Dayan, P. (2000). A model of hippocampally dependent navigation, using the temporal difference learning rule. *Hippocampus*, 10:1–16.

Fox, E., Sudderth, E., Jordan, M., and Willsky, A. (2011). Bayesian nonparametric inference of switching dynamic linear models. *Signal Processing, IEEE Transactions on*, 59(4):1569–1585.

Frank, M., Goldwater, S., Griffiths, T., and Tenenbaum, J. (2010). Modeling human performance in statistical word segmentation. *Cognition*, 117(2):107–125.

Frankland, P., Ding, H., Takahashi, E., Suzuki, A., Kida, S., and Silva, A. (2006). Stability of recent and remote contextual fear memory. *Learning & Memory*, 13(4):451–457.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456):815.

Frohardt, R. J., Guarraci, F. A., and Bouton, M. E. (2000). The effects of neurotoxic hippocampal lesions on two effects of context after fear extinction. *Behavioral Neuroscience*, 114:227–240.

Fuchs, R., Bell, G., Ramirez, D., Eaddy, J., and Su, Z. (2009). Basolateral amygdala involvement in memory reconsolidation processes that facilitate drug context-induced cocaine seeking. *European Journal of Neuroscience*, 30(5):889–900.

Fuhs, M. C. and Touretzky, D. S. (2007). Context Learning in the Rodent Hippocampus. *Neural Computation*, 19:3173–3215.

Galluccio, L. and Rovee-Collier, C. (2005). Updating reactivated memories in infancy: Ii. time passage and repetition effects. *Developmental psychobiology*, 47(1):18–30.

Geisler, W. and Diehl, R. (2003). A bayesian approach to the evolution of perceptual and cognitive systems. *Cognitive Science*, 27(3):379–402.

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Gershman, S. and Blei, D. (2012). A tutorial on Bayesian nonparametric models. *Journal of Mathematical Psychology*, 56:1–12.

Gershman, S., Blei, D., and Niv, Y. (2010). Context, learning, and extinction. *Psychological Review*, 117(1):197–209.

Gershman, S., Moore, C., Todd, M., Norman, K., and Sederberg, P. (2012). The successor representation and temporal context. *Neural Computation*, 24:1–16.

Gershman, S. and Niv, Y. (2010). Learning latent structure: carving nature at its joints. *Current Opinion in Neurobiology*.

Glimcher, P. (2011). Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences*, 108(Supplement 3):15647–15654.

Gluck, M. and Myers, C. (1993). Hippocampal mediation of stimulus representation: A computational theory. *Hippocampus*, 3(4):491–516.

Gold, P. and King, R. (1974). Retrograde amnesia: Storage failure versus retrieval failure. *Psychological Review*, 81:465–469.

Goldstone, R. (1995). Effects of categorization on color perception. *Psychological Science*, pages 298–304.

Goldwater, S., Griffiths, T., and Johnson, M. (2009). A bayesian framework for word segmentation: Exploring the effects of context. *Cognition*, 112(1):21–54.

Griffiths, T., Chater, N., Kemp, C., Perfors, A., and Tenenbaum, J. (2010). Probabilistic models of cognition: exploring representations and inductive biases. *Trends in Cognitive Sciences*, 14(8):357–364.

Grossberg, S. (1976). Adaptive pattern classification and universal recoding: I. Parallel development and coding of neural feature detectors. *Biological cybernetics*, 23(3):121–134.

Gureckis, T. and Goldstone, R. (2008). The effect of the internal structure of categories on perception. In *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, pages 1876–1881. Citeseer.

Hall, G. and Honey, R. C. (1989). Contextual effects in conditioning, latent inhibition, and habituation: Associative and retrieval functions of contextual cues. *Journal of Experimental Psychology: Animal Behavior Processes*, 15:232–241.

Hasselmo, M. E., Bodelón, C., and Wyble, B. (2002). A proposed function for hippocampal theta rhythm: separate phases of encoding and retrieval enhance reversal of prior learning. *Neural Computation*, 14(4):793–817.

Hasselmo, M. E. and Eichenbaum, H. (2005). Hippocampal mechanisms for the context-dependent retrieval of episodes. *Neural Networks*, 18:1172–1190.

Hemmer, P. and Steyvers, M. (2009). A Bayesian Account of Reconstructive Memory. *Topics in Cognitive Science*, 1(1):189–202.

Hernandez, P. and Kelley, A. (2004). Long-term memory for instrumental responses does not undergo protein synthesis-dependent reconsolidation upon retrieval. *Learning & memory*, 11(6):748–754.

Hertz, J., Krogh, A., and Palmer, R. (1991). *Introduction to the Theory of Neural Computation*. Addison Wesley Publishing Company, Boston, MA.

Hinderliter, C., Webster, T., and Riccio, D. (1975). Amnesia induced by hypothermia as a function of treatment-test interval and recooling in rats. *Animmal Learning & Behavior*, 3:257–263.

Hintzman, D. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, 95(4):528–551.

Hirsh, R. (1974). The hippocampus and contextual retrieval of information from memory: A theory. *Behavioral Biology*, 12(4):421.

Holland, M. and Lockhead, G. (1968). Sequential effects in absolute judgments of loudness. *Attention, Perception, & Psychophysics*, 3(6):409–414.

Holland, P. C. (1988). Excitation and inhibition in unblocking. *Journal of Experimental Psychology: Animal Behavioral Processes*, 14:261–279.

Honey, R. C. and Good, M. (1993). Selective hippocampal lesions abolish the contextual specificity of latent inhibition and conditioning. *Behavioral Neuroscience*, 107:23–33.

Hopfield, J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, 79:2554–2558.

Howard, M. and Kahana, M. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, 46(3):269–299.

Hupbach, A., Gomez, R., Hardt, O., and Nadel, L. (2007). Reconsolidation of episodic memories: A subtle reminder triggers integration of new information. *Learning & Memory*, 14(1-2):47–53.

Hupbach, A., Gomez, R., and Nadel, L. (2011). Episodic memory updating: The role of context familiarity. *Psychonomic Bulletin & Review*, pages 1–11.

Hupbach, A., Hardt, O., Gomez, R., and Nadel, L. (2008). The dynamics of memory: Context-dependent updating. *Learning & Memory*, 15(8):574–579.

Hupbach, D., Gomez, R., and Nadel, L. (2009). Episodic memory reconsolidation: Updating or source confusion? *Memory*, 17(5):502–510.

Huttenlocher, J., Hedges, L., and Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review*, 98(3):352–376.

Huttenlocher, J., Hedges, L., and Vevea, J. (2000). Why do categories affect stimulus judgment?. *Journal of Experimental Psychology: General*, 129(2):220–241.

Izard, V. and Dehaene, S. (2008). Calibrating the mental number line. *Cognition*, 106(3):1221–1247.

James, W. (1901). *Talks to Teachers on Psychology*. H. Holt and Co.

Jarome, T., Kwapis, J., Werner, C., Parsons, R., Gafford, G., and Helmstetter, F. (2012). The timing of multiple retrieval events can alter glur1 phosphorylation and the requirement for protein synthesis in fear memory reconsolidation. *Learning & Memory*, 19(7):300–306.

Jarrard, L. (1993). On the role of the hippocampus in learning and memory in the rat. *Behavioral and Neural Biology*, 60(1):9–26.

Jaynes, E. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge, UK.

Jenkins, J. and Dallenbach, K. (1924). Obliviscence during sleep and waking. *The American Journal of Psychology*, 35(4):605–612.

Ji, J. and Maren, S. (2005). Electrolytic lesions of the dorsal hippocampus disrupt renewal of conditional fear after extinction. *Learning and Memory*, 12:270–276.

Ji, J. and Maren, S. (2007). Hippocampal involvement in contextual modulation of fear extinction. *Hippocampus*, 17:749–58.

Johnson, A., van der Meer, M. A., and Redish, A. D. (2007). Integrating hippocampus and striatum in decision-making. *Current Opinion in Neurobiology*, 17:692–697.

Johnson, J., McDuff, S., Rugg, M., and Norman, K. (2009). Recollection, familiarity, and cortical reinstatement: a multivoxel pattern analysis. *Neuron*, 63(5):697–708.

Jones, B., Bukoski, E., Nadel, L., and Fellous, J. (2012). Remaking memories: Reconsolidation updates positively motivated spatial memory in rats. *Learning & Memory*, 19(3):91–98.

Jones, M. and Love, B. (2011). Bayesian fundamentalism or enlightenment? on the explanatory status and theoretical contributions of bayesian models of cognition. *Behavioral and Brain Sciences*, 34(04):169–188.

Kakade, S. and Dayan, P. (2002). Acquisition and extinction in autoshaping. *Psychological Review*, 109:2002.

Kalman, R. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45.

Kass, R. and Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430).

Kehoe, E. and White, N. (2002). Extinction revisited: Similarities between extinction and reductions in us intensity in classical conditioning of the rabbits nictitating membrane response. *Learning & Behavior*, 30(2):96–111.

Kemp, C., Tenenbaum, J., Niyogi, S., and Griffiths, T. (2010). A probabilistic model of theory formation. *Cognition*, 114(2):165–196.

Kemp, C. and Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, 105:10687–10692.

Kindt, M. and Soeter, M. (2011). Reconsolidation in a human fear conditioning study: A test of extinction as updating mechanism. *Biological Psychology*.

Koob, G. and Volkow, N. (2009). Neurocircuitry of addiction. *Neuropsychopharmacology*, 35(1):217–238.

Kopp, R., Bohdanecky, Z., and Jarvik, M. (1966). Long temporal gradient of retrograde amnesia for a well-discriminated stimulus. *Science*, 153(3743):1547.

Kruschke, J. (1992). Alcove: An exemplar-based connectionist model of category learning. *Psychological Review*, 99(1):22–44.

Kruschke, J. (2008). Bayesian approaches to associative learning: From passive to active learning. *Learning & Behavior*, 36(3):210–226.

Kuhl, B., Rissman, J., Chun, M., and Wagner, A. (2011). Fidelity of neural reactivation reveals competition between memories. *Proceedings of the National Academy of Sciences*, 108(14):5903.

Larrauri, J. A. and Schmajuk, N. A. (2008). Attentional, associative, and configural mechanisms in extinction. *Psychological Review*, 115:640–676.

Lattal, K. and Abel, T. (2004). Behavioral impairments caused by injections of the protein synthesis inhibitor anisomycin after contextual retrieval reverse with time. *Proceedings of the National Academy of Sciences of the United States of America*, 101(13):4667.

Lee, J., Milton, A., and Everitt, B. (2006). Reconsolidation and extinction of conditioned fear: inhibition and potentiation. *The Journal of neuroscience*, 26(39):10051–10056.

Lee, S., Choi, J., Lee, N., Lee, H., Kim, J., Yu, N., Choi, S., Lee, S., Kim, H., and Kaang, B. (2008). Synaptic protein degradation underlies destabilization of retrieved fear memory. *Science*, 319(5867):1253–1256.

Leutgeb, J., Leutgeb, S., Treves, A., Meyer, R., Barnes, C., McNaughton, B., Moser, M., and Moser, E. (2005). Progressive transformation of hippocampal neuronal representations in morphed environments. *Neuron*, 48(2):345–358.

Levy, W., Hocking, A., and Wu, X. (2005). Interpreting hippocampal function as recoding and forecasting. *Neural Networks*, 18(9):1242–1264.

Lewis, D., Misanin, J., and Miller, R. (1968). Recovery of memory following amnesia. *Nature*, 220:704–705.

Li, M. and Vitanyi, P. (2008). *An Introduction to Kolmogorov Complexity and its Applications*. Springer-Verlag.

Liberman, A., Harris, K., Hoffman, H., and Griffith, B. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, 54(5):358–368.

Lisman, J. E. and Grace, A. A. (2005). The hippocampal-VTA loop: controlling the entry of information into long-term memory. *Neuron*, 46:703–713.

Love, B., Medin, D., and Gureckis, T. (2004). SUSTAIN: A network model of category learning. *Psychological Review*, 111:309–332.

Love, B. C. and Gureckis, T. M. (2007). Models in search of a brain. *Cognitive, Affective and Behavioral Neuroscience*, 7:90–108.

Ma, X., Zhang, J., and Yu, L. (2011). Post-retrieval extinction training enhances or hinders the extinction of morphine-induced conditioned place preference in rats dependent on the retrieval-extinction interval. *Psychopharmacology*, pages 1–8.

Mactutus, C., Riccio, D., and Ferek, J. (1979). Retrograde amnesia for old (reactivated) memory: some anomalous characteristics. *Science*, 204(4399):1319.

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

Marr, D. (1982). *Vision*. "W. H. Freeman and Company, San Francisco, CA.

Martin, P. D. and Berthoz, A. (2002). Development of spatial firing in the hippocampus of young rats. *Hippocampus*, 12:465–80.

McClelland, J. and Rumelhart, D. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General*, 114(2):159–188.

McClelland, J. L., McNaughton, B. L., and 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.

McGaugh, J. (2000). Memory–a century of consolidation. *Science*, 287(5451):248–251.

McGeoch, J. (1932). Forgetting and the law of disuse. *Psychological Review*, 39(4):352–370.

McNally, G., Johansen, J., and Blair, H. (2011). Placing prediction into the fear circuit. *Trends in Neurosciences*, 34:283–292.

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

Medin, D. and Schaffer, M. (1978). Context theory of classification learning. *Psychological Review*, 85(3):207–238.

Milekic, M. and Alberini, C. (2002). Temporally graded requirement for protein synthesis following memory reactivation. *Neuron*, 36(3):521–525.

Miller, R. and Laborda, M. (2011). Preventing recovery from extinction and relapse. *Current Directions in Psychological Science*, 20(5):325–329.

Miller, R. and Matzel, L. (2006). Retrieval failure versus memory loss in experimental amnesia: Definitions and processes. *Learning & Memory*, 13(5):491–497.

Milton, A., Lee, J., Butler, V., Gardner, R., and Everitt, B. (2008). Intra-amygdala and systemic antagonism of nmda receptors prevents the reconsolidation of drug-associated memory and impairs subsequently both novel and previously acquired drug-seeking behaviors. *The Journal of Neuroscience*, 28(33):8230–8237.

Misanin, J., Miller, R., and Lewis, D. (1968). Retrograde amnesia produced by electroconvulsive shock after reactivation of a consolidated memory trace. *Science*, 160(3827):554.

Monfils, M., Cowansage, K., Klann, E., and LeDoux, J. (2009). Extinction-reconsolidation boundaries: key to persistent attenuation of fear memories. *Science*, 324(5929):951.

Morris, R., Inglis, J., Ainge, J., Olverman, H., Tulloch, J., Dudai, Y., and Kelly, P. (2006). Memory reconsolidation: sensitivity of spatial memory to inhibition of protein synthesis in dorsal hippocampus during encoding and retrieval. *Neuron*, 50(3):479–489.

Muller, G. and Pilzecker, A. (1900). Experimentelle bietrage zur lehre bom gedachtnisses. *Zeits. Fr Psych.*, 1:1–288.

Nadel, L., Winocur, G., Ryan, L., and Moscovitch, M. (2007). Systems consolidation and hippocampus: two views. *Debates in Neuroscience*, 1(2):55–66.

Nader, K. and Hardt, O. (2009). A single standard for memory: the case for reconsolidation. *Nature Reviews Neuroscience*, 10(3):224–234.

Nader, K., Schafe, G., and Le Doux, J. (2000). Fear memories require protein synthesis in the amygdala for reconsolidation after retrieval. *Nature*, 406(6797):722–726.

Napier, R., Macrae, M., and Kehoe, E. (1992). Rapid reacquisition in conditioning of the rabbit's nictitating membrane response. *Journal of Experimental Psychology: Animal Behavior Processes*, 18(2):182–192.

Nassar, M., Wilson, R., Heasly, B., and Gold, J. (2010). An approximately bayesian delta-rule model explains the dynamics of belief updating in a changing environment. *The Journal of Neuroscience*, 30(37):12366–12378.

Neal, R. and Hinton, G. (1998). A view of the em algorithm that justifies incremental, sparse, and other variants. In Jordan, M., editor, *Learning in Graphical Models*, pages 355–368. MIT Press.

Norman, K., Detre, G., and Polyn, S. (2006). Computational models of episodic memory. In Sun, R., editor, *The Cambridge Handbook of Computational Psychology*, pages 189–224. Cambridge University Press.

Nosofsky, R. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, 115(1):39–57.

Nosofsky, R. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14(4):700–708.

Nosofsky, R. (1991). Relation between the rational model and the context model of categorization. *Psychological Science*, 2(6):416–421.

Nyberg, L. (2005). Any novelty in hippocampal formation and memory? *Current Opinion in Neurology*, 18(4):424.

O'Keefe, J. and Nadel, L. (1978). *The Hippocampus as a Cognitive Map*. Oxford University Press, USA.

O'Reilly, R. C. and McClelland, J. L. (1994). Hippocampal conjunctive encoding, storage, and recall: avoiding a trade-off. *Hippocampus*, 4:661–682.

Osan, R., Tort, A., and Amaral, O. (2011). A mismatch-based model for memory reconsolidation and extinction in attractor networks. *PloS One*, 6(8):e23113.

Oyarzún, J., Lopez-Barroso, D., Fuentemilla, L., Cucurell, D., Pedraza, C., Rodriguez-Fornells, A., and de Diego-Balaguer, R. (2012). Updating fearful memories with extinction training during reconsolidation: A human study using auditory aversive stimuli. *PloS one*, 7(6):e38849.

Paller, K. and Wagner, A. (2002). Observing the transformation of experience into memory. *Trends in Cognitive Sciences*, 6(2):93–102.

Pavlov, I. (1927). *Conditioned Reflexes*. Oxford University Press.

Pearce, J. M. and Bouton, M. E. (2001). Theories of associative learning in animals. *Annual Review of Psychology*, 52:111–139.

Pedreira, M. and Maldonado, H. (2003). Protein synthesis subserves reconsolidation or extinction depending on reminder duration. *Neuron*, 38(6):863–869.

Pedreira, M., Pérez-Cuesta, L., and Maldonado, H. (2004). Mismatch between what is expected and what actually occurs triggers memory reconsolidation or extinction. *Learning & Memory*, 11(5):579.

Pezze, M. and Feldon, J. (2004). Mesolimbic dopaminergic pathways in fear conditioning. *Progress in neurobiology*, 74(5):301–320.

Pitman, J. (2002). *Combinatorial Stochastic Processes*. Notes for Saint Flour Summer School. Techincal Report 621, Dept. Statistics, UC Berkeley.

Polyn, S. and Kahana, M. (2008). Memory search and the neural representation of context. *Trends in Cognitive Sciences*, 12(1):24–30.

Polyn, S., Natu, V., Cohen, J., and Norman, K. (2005). Category-specific cortical activity precedes retrieval during memory search. *Science*, 310(5756):1963–1966.

Pothos, E. and Chater, N. (2002). A simplicity principle in unsupervised human categorization. *Cognitive Science*, 26(3):303–343.

Power, A., Berlau, D., McGaugh, J., and Steward, O. (2006). Anisomycin infused into the hippocampus fails to block "reconsolidation" but impairs extinction: The role of re-exposure duration. *Learning & Memory*, 13(1):27.

Preminger, S., Blumenfeld, B., Sagi, D., and Tsodyks, M. (2009). Mapping dynamic memories of gradually changing objects. *Proceedings of the National Academy of Sciences*, 106(13):5371–5376.

Preminger, S., Sagi, D., and Tsodyks, M. (2007). The effects of perceptual history on memory of visual objects. *Vision research*, 47(7):965–973.

Quartermain, D., Paolino, R., and Miller, N. (1965). A brief temporal gradient of retrograde amnesia independent of situational change. *Science*, 149(3688):1116.

Quirk, G. J., Garcia, R., and Gonzalez-Lima, F. (2006). Prefrontal mechanisms in extinction of conditioned fear. *Biological Psychiatry*, 60:337–343.

Raaijmakers, J. and Shiffrin, R. (1981). Search of associative memory. *Psychological Review*, 88:93.

Raaijmakers, J. and Shiffrin, R. (1992). Models for recall and recognition. *Annual Review of Psychology*, 43:205–234.

Rasmussen, C. (2000). The infinite Gaussian mixture model. *Advances in Neural Information Processing Systems*, 12:554–560.

Ratcliff, R., Clark, S., and Shiffrin, R. (1990). List-strength effect: I. data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(2):163–178.

Redish, A., Jensen, S., Johnson, A., and Kurth-Nelson, Z. (2007). Reconciling reinforcement learning models with behavioral extinction and renewal: Implications for addiction, relapse, and problem gambling. *Psychological Review*, 114(3):784–805.

Rescorla, R. (2004). Spontaneous recovery. *Learning & Memory*, 11(5):501.

Rescorla, R. and Heth, C. (1975). Reinstatement of fear to an extinguished conditioned stimulus. *Journal of Experimental Psychology: Animal Behavior Processes*, 1(1):88–96.

Rescorla, R. A. and Wagner, A. R. (1972). A theory of of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In Black, A. and Prokasy, W., editors, *Classical Conditioning II: Current Research and theory*, pages 64–99. Appleton-Century-Crofts, New York, NY.

Riccio, D., Millin, P., and Bogart, A. (2006). Reconsolidation: A brief history, a retrieval view, and some recent issues. *Learning & Memory*, 13(5):536.

Ricker, S. and Bouton, M. (1996). Reacquisition following extinction in appetitive conditioning. *Animal Learning and Behavior*, 24(4):423–436.

Robert, C. and Casella (2004). *Monte Carlo Statistical Methods*. Springer New York.

Rohrbaugh, M. and Riccio, D. (1970). Paradoxical enhancement of learned fear. *Journal of Abnormal Psychology*, 75(2):210–216.

Rolls, E. (2010). A computational theory of episodic memory formation in the hippocampus. *Behavioural Brain Research*, 215(2):180–196.

Rudy, J. W. (1993). Contextual conditioning and auditory cue conditioning dissociate during development. *Behavioral Neuroscience*, 107:887–91.

Rumpel, S., LeDoux, J., Zador, A., and Malinow, R. (2005). Postsynaptic receptor trafficking underlying a form of associative learning. *Science*, 308(5718):83–88.

Sailor, K. and Antoine, M. (2005). Is memory for stimulus magnitude bayesian? *Memory & Cognition*, 33(5):840–851.

Sanborn, A., Griffiths, T., and Navarro, D. (2006). A more rational model of categorization. In *Proceedings of the 28th annual conference of the cognitive science society*, pages 726–731.

Sanborn, A., Griffiths, T., and Navarro, D. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, 117(4):1144–1167.

Schafe, G. and LeDoux, J. (2000). Memory consolidation of auditory pavlovian fear conditioning requires protein synthesis and protein kinase A in the amygdala. *Journal of Neuroscience*, 20(18):96.

Schiller, D., Monfils, M., Raio, C., Johnson, D., LeDoux, J., and Phelps, E. (2010). Preventing the return of fear in humans using reconsolidation update mechanisms. *Nature*, 463(7277):49–53.

Schmajuk, N. A., Buhusi, C., and Gray, J. A. (1996). An attentional-configural model of classical conditioning. *Journal of Mathematical Psychology*, 40:358–358.

Schultz, W., Dayan, P., and Montague, P. R. (1997). A neural substrate of prediction and reward. *Science*, 275:1593–1599.

Sederberg, P., Gershman, S., Polyn, S., and Norman, K. (2011). Human memory

reconsolidation can be explained using the temporal context model. *Psychonomic Bulletin & Review*, 18:455–468.

Sederberg, P., Howard, M., and Kahana, M. (2008). A context-based theory of recency and contiguity in free recall. *Psychological Review*, 115(4):893–912.

Shiffrin, R., Ratcliff, R., and Clark, S. (1990). List-strength effect: Ii. theoretical mechanisms. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 16(2):179–195.

Socher, R., Gershman, S., Perotte, A., Sederberg, P., Blei, D., and Norman, K. (2009). A bayesian analysis of dynamics in free recall. In *Neural Information Processing Systems*.

Spear, N. (1973). Retrieval of memory in animals. *Psychological Review*, 80(3):163–194.

Squire, L. (2006). Lost forever or temporarily misplaced? the long debate about the nature of memory impairment. *Learning & Memory*, 13(5):522–529.

Squire, L. and Alvarez, P. (1995). Retrograde amnesia and memory consolidation: a neurobiological perspective. *Current Opinion in Neurobiology*, 5(2):169–177.

Steyvers, M., Griffiths, T., and Dennis, S. (2006). Probabilistic inference in human semantic memory. *Trends in Cognitive Sciences*, 10(7):327–334.

Summerfield, C., Behrens, T., and Koechlin, E. (2011). Perceptual classification in a rapidly changing environment. *Neuron*, 71(4):725–736.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA.

Suzuki, A., Josselyn, S., Frankland, P., Masushige, S., Silva, A., and Kida, S. (2004). Memory reconsolidation and extinction have distinct temporal and biochemical signatures. *Journal of Neuroscience*, 24(20):4787.

Vinogradova, O. (2001). Hippocampus as comparator: role of the two input and two output systems of the hippocampus in selection and registration of information. *Hippocampus*, 11(5):578–598.

von Restorff, H. (1933). Uber die wirkung von bereichsbildungen im spurenfeld. *Psychologische Forschung*, 18:299–342.

Wallace, W. (1965). Review of the historical, empirical, and theoretical status of the von restorff phenomenon. *Psychological Bulletin*, 63(6):410.

Wallis, G. and Bülthoff, H. (2001). Effects of temporal association on recognition memory. *Proceedings of the National Academy of Sciences*, 98(8):4800–4804.

Wang, L. and Dunson, D. (2011). Fast bayesian inference in dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, 20(1):196–216.

Wang, S., de Oliveira Alvares, L., and Nader, K. (2009). Cellular and systems mechanisms of memory strength as a constraint on auditory fear reconsolidation. *Nature Neuroscience*, 12(7):905–912.

Widrow, B. and Hoff, M. (1960). Adaptive switching circuits.

Wills, T., Lever, C., Cacucci, F., Burgess, N., and O'Keefe, J. (2005). Attractor dynamics in the hippocampal representation of the local environment. *Science*, 308(5723):873–876.

Wilson, A., Brooks, D. C., and Bouton, M. E. (1995). The role of the rat hippocampal system in several effects of context in extinction. *Behavioral Neuroscience*, 109:828–836.

Wilson, R. and Finkel, L. (2009). A neural implementation of the Kalman filter. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C. K. I., and Culotta, A., editors, *Advances in Neural Information Processing Systems 22*, pages 2062–2070.

Winters, B., Tucci, M., and DaCosta-Furtado, M. (2009). Older and stronger object memories are selectively destabilized by reactivation in the presence of new information. *Learning & Memory*, 16(9):545.

Wixted, J. (2004). The psychology and neuroscience of forgetting. *Annual Review of Psychology*, 55(1):235.

Woods, A. and Bouton, M. (2007). Occasional reinforced responses during extinction can slow the rate of reacquisition of an operant response. *Learning & Motivation*, 38(1):56–74.

Yap, C. S. and Richardson, R. (2005). Latent inhibition in the developing rat: an examination of context-specific effects. *Developmental Psychobiology*, 47:55–65.

Yap, C. S. and Richardson, R. (2007). Extinction in the developing rat: an examination of renewal effects. *Developmental Psychobiology*, 49:565–575.

Yehuda, R. and LeDoux, J. (2007). Response variation following trauma: a translational neuroscience approach to understanding PTSD. *Neuron*, 56(1):19–32.

Yu, A. and Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, 46(4):681–692.

Zeithamova, D. and Maddox, W. (2009). Learning Mode and Exemplar Sequencing in Unsupervised Category Learning. *Learning, Memory*, 35(3):731–741.

Zhu, X., Ghahramani, Z., and Lafferty, J. (2005). Time-sensitive dirichlet process mixture models. *Technical Report, CMU-CALD-05-104, Carnegie Mellon University.*