Running title: THEORY OF RECONSOLIDATION

The Computational Nature of Memory Reconsolidation

Samuel J. Gershman[1], Marie-H Monfils[2], Kenneth A. Norman[3], and Yael Niv[3]
[1]Department of Psychology and Center for Brain Science, Harvard University
[2]Department of Psychology, University of Texas at Austin
[3]Princeton Neuroscience Institute and Department of Psychology, Princeton University

Address for correspondence:
Samuel Gershman
Department of Psychology
Harvard University
52 Oxford St., room 295.05
Cambridge, MA 02138
Phone: 773-607-9817
E-mail: gershman@fas.harvard.edu

Abstract

Retrieval can render memories labile, allowing them to be modified or erased by behavioral or pharmacological intervention. This phenomenon, known as reconsolidation, defies explanation in terms of classical associative learning theories, prompting a reconsideration of basic learning mechanisms in the brain. We propose that two mechanisms interact to produce reconsolidation: an associative learning mechanism of the form posited by classical theories, and a structure learning mechanism that discovers the units of association by segmenting the stream of experience into statistically distinct clusters (latent causes). We derive this framework from statistical principles, and present a mechanistic implementation. Simulations demonstrate that it can reproduce the major experimental findings from studies of reconsolidation, including dependence on the strength and age of memories, the interval between memory retrieval and extinction, and prediction errors following retrieval. In addition, we present new experimental data confirming the theory's prediction that performing part of extinction prior to retrieval attenuates reconsolidation.

KEYWORDS: memory, Bayesian, extinction, fear conditioning

Introduction

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 & Pilzecker, 1900). This phenomenon raises a basic question about memory: Once consolidated, can traces ever be modified again?

Answers 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, Miller, & Lewis, 1968). Using a Pavlovian fear conditioning task, Misanin et al. (1968) found that electroconvulsive shock had no effect on a fear memory acquired a day previously; however, if the animal was briefly reexposed to the acquisition 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, Millin, & Bogart, 2006, for a historical overview).

Contemporary neuroscientific interest in reconsolidation was ignited by Nader, Schafe, and Le Doux (2000), who showed that retrograde amnesia for an acquired fear memory could be produced by injection of a protein synthesis inhibitor (PSI) into the lateral nucleus of the amygdala shortly after reexposure to the acquisition cue. Subsequent studies have generated a detailed neural and behavioral characterization of reconsolidation, including a large cast of molecular mechanisms (Tronson & Taylor, 2007), and a number of boundary conditions on the occurrence of reconsolidation (Duvarci & Nader, 2004; Nader & Hardt, 2009). Moreover, there is now evidence that memory updating can be obtained with a purely behavioral procedure (Monfils, Cowansage, Klann, & LeDoux, 2009; Schiller et al., 2010). These findings have lead to the view that the function of reconsolidation is to allow memories to be updated by new information (Alberini, 2007; J. Lee, 2009), but many basic mechanistic questions remain ambiguous or unanswered (Squire, 2006).

Reconsolidation challenges most existing computational models of Pavlovian conditioning. For concreteness, we focus on the most well-known of these, the Rescorla-Wagner model (Rescorla & Wagner, 1972). This model posits that, over the course of acquisition, the animal learns an association between the conditional stimulus (CS, e.g., tone) and the unconditional stimulus (US, e.g., shock). The main weakness of the Rescorla-Wagner model is its prediction that presenting the CS repeatedly by itself (extinction) will erase the cue-outcome association formed during acquisition—in other words, the model predicts that *extinction is unlearning*. It is now widely accepted that this assumption, shared by a large class of models, is wrong (Delamater, 2004; Gallistel, 2012).

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 (Bouton & Bolles, 1979b; Pavlov, 1927; Rescorla & Heth, 1975). Conditioned responding can also be recovered if the animal is returned to the acquisition context, a phenomenon known as *renewal* (Bouton & Bolles, 1979a).

Bouton (1993) interpreted the attenuation of responding after extinction in terms of the formation of an extinction memory that competes for retrieval with the acquisition memory; this retroactive interference can be relieved by a change in temporal context or the presence of retrieval cues, thereby leading to recovery (see also Miller & Laborda, 2011). Central to retrieval-based accounts is the idea that the associations learned during acquisition are linked to the spatiotemporal context of the acquisition session, and as a result they are largely unaffected by extinction. 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 acquisition context (e.g., renewal, reinstatement) or attenuate elements of the extinction context (e.g., spontaneous recovery). According to this view, reconsolidation (i.e., modification of the original memory) should occur when the acquisition and extinction phases are linked to the same spatiotemporal context. 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 paper, we present a computational theory of Pavlovian conditioning that attempts to answer these questions, and use it to understand memory reconsolidation. We then show how this model can account for a wide variety of reconsolidation phenomena, including fine-grained temporal dynamics.

Like the Rescorla-Wagner model (Figure 1A), our theory posits the learning of CS-US associations; however, in our model 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, Daw, & Touretzky, 2006; Gershman, Blei, & Niv, 2010; Gershman, Jones, Norman, Monfils, & Niv, 2013; Gershman & Niv, 2012; Gershman, Norman, & Niv, 2015; Soto, Gershman, & Niv, 2014). We 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 our theory, the animal learns a different set of associations for each latent cause, flexibly inferring new causes when existing causes no longer predict the currently observed CS-US relationship accurately (Figure 1B). This allows the theory to move beyond the "extinction=unlearning" assumption by assuming that different latent causes are inferred during acquisition and extinction, thus two different associations are learned (see also Redish, Jensen, Johnson, & Kurth-Nelson, 2007).

According to our theory, reconsolidation arises when CS reexposure provides evidence to the

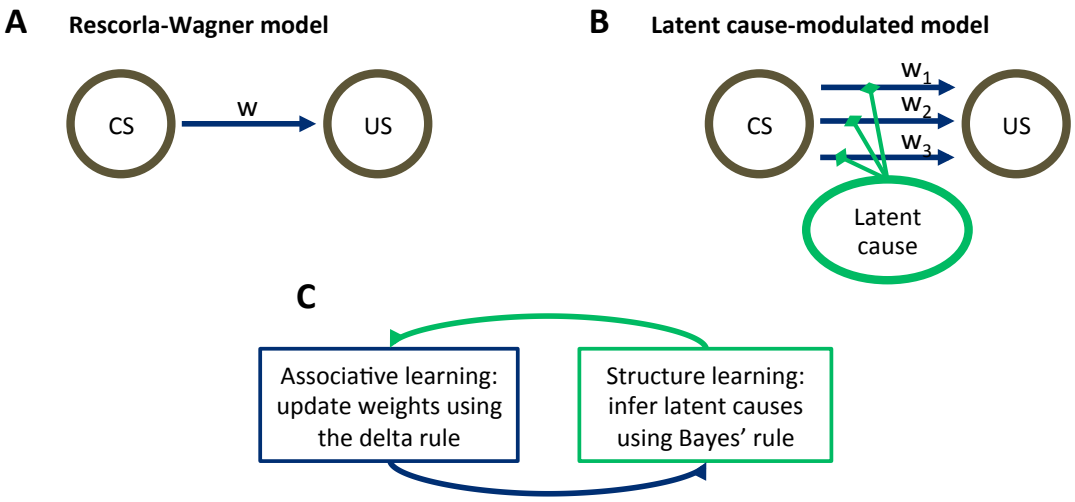

Figure 1: **Model schematic**. (A) The associative structure underlying the Rescorla-Wagner model. The associative strength between a conditioned stimulus (CS) and an unconditioned stimulus (US) is encoded by a scalar weight, w. (B) The associative structure underlying the latent cause-modulated model. 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.

animal that the latent cause assigned to the acquisition phase is once again active, making that cause's associations eligible for updating (or disruption by PSIs). We show that this theory is able to account for the main boundary conditions on reconsolidation using PSIs (Nader & 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 we confirm experimentally.

A Rational Analysis of Pavlovian Conditioning

Our theory is derived from a rational analysis (cf. Anderson, 1990) of the learning problem facing an animal in Pavlovian conditioning. A rational analysis begins with a hypothetical internal model that describes how latent causes give rise to observed stimuli—a "generative model" that embodies the animal's beliefs about causes and effects in the world. The task of structure learning, according to our analysis, is to "invert" the internal model, using

the observed stimuli to make inferences about the the latent causes that generated them (Gershman & Niv, 2010). The optimal inversion of the generative process is stipulated by Bayes' rule, a basic law of probability theory. 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 was likely to have generated the current sensory inputs.

High-level description of the theory

Before elaborating the technical details of our theory, we first provide a high-level description. The basic computational framework consists of two interacting sub-systems (Figure 1C): an associative learning system updates a set of CS-US associations using a delta rule (Rescorla & Wagner, 1972; Sutton & Barto, 1998; Widrow & Hoff, 1960), 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 almost identical to the Rescorla-Wagner model, with the key difference that the system can maintain multiple sets of associations (one for each latent cause; Figure 1B) 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).

To infer the posterior over latent causes, the structure learning system makes certain assumptions about the statistics of latent causes. Informally, the main assumptions we 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 & Vitányi, 2003; Gershman & Niv, 2013). We use 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 (Gershman & Blei, 2012).

- *Contiguity principle*: the closer two observations occur in time, the more likely it is that 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. We now describe the theory in greater technical detail.

The internal model

Our specification of the animal's internal model consists of three parts: (1) a distribution over latent causes, (2) a conditional distribution over CS configurations given latent causes, and (3) a conditional distribution over the US given the CS configuration. We 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 D-dimensional CS configuration at time t. The distribution over r_t and $\mathbf{x}_t$ is determined by a latent cause z_t. Specifically, the CS configuration is drawn from a Gaussian distribution:

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

where μ_{kd} is the expected intensity of the dth CS given cause k is active, and σ_x^2 is its variance. We 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 & Dayan, 2002; Kruschke, 2008), we assume that the US is generated by a weighted combination of the CSs corrupted by Gaussian noise:

$$P(r_t|z_t = k) = \mathcal{N}\left(r_t; \textstyle\sum_{d=1}^{D} w_{kd} x_{td}, \sigma_r^2\right). \tag{2}$$

According to the animal's internal model, on each trial the active latent cause z_t is drawn from the following distribution:

$$P(z_t = k|\mathbf{z}_{1:t-1}) \propto \begin{cases} \sum_{t'<t} \mathcal{K}(\tau(t) - \tau(t'))\mathbb{I}[z_{t'} = k] & \text{if } k \leq K \text{ (i.e., } k \text{ is an old cause)} \\ \alpha & \text{otherwise (i.e., } k \text{ is a new cause)} \end{cases} \tag{3}$$

where $\mathbb{I}[\cdot] = 1$ when its arguments are true (0 otherwise), $\tau(t)$ is the time at which trial t occurred and $\mathcal{K}$ is a temporal kernel that governs the temporal dependence between latent causes. Intuitively, this means that the CS configuration on a particular trial will likely be generated by the same latent cause as was active on other trials that occurred nearby in time. The "concentration" parameter α determines the bias towards generating a completely new latent cause. This infinite-capacity distribution over latent causes imposes the simplicity principle described in the previous section—a small number of latent causes, each active for a continuous period of time, is more likely *a priori* than a large number of intertwined causes. The distribution defined by Eq. 3 was first introduced by Zhu, Ghahramani, and Lafferty (2005) in their "time-sensitive" generalization of the Chinese restaurant process (Aldous, 1985).[1] Variants of this distribution have been widely used in cognitive science to model probabilistic reasoning about combinatorial objects of unbounded cardinality (e.g., Anderson, 1991; Collins & Koechlin, 2012; Gershman et al., 2010; Goldwater, Griffiths, &

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

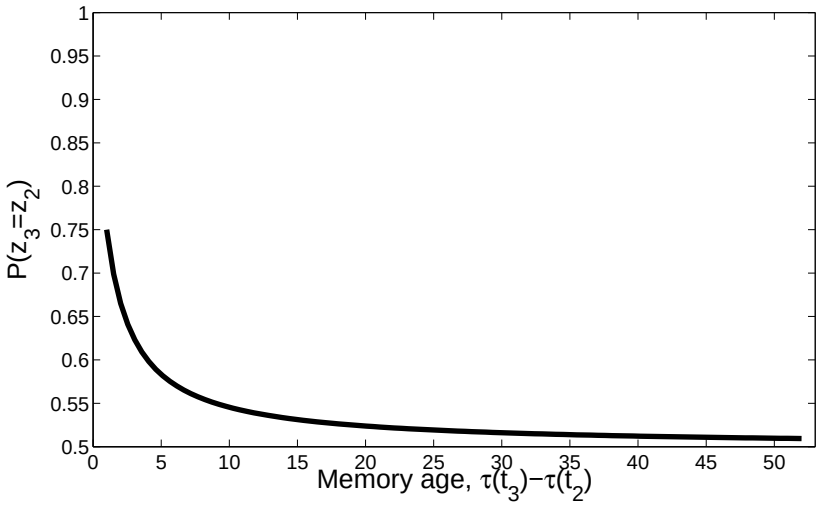

Figure 2: **Temporal compression with the power law kernel**. We assume that t_1 was generated by cause z_1, two timepoints later t_2 was generated by cause z_2, and a variable number of timepoints later t_3 was generated by cause z_3. To illustrate the time compression property we have assumed that the probability of a new cause is 0 (i.e., $\alpha = 0$) so inference at t_3 is constrained to one of the previous causes. The probability of trial 3 being generated by one of two previous causes diminishes as the temporal distance between $\tau(t_3)$ and the time of the previous trial $\tau(t_2)$ increases, that is, as the memory for t_2 recedes into the past.

Johnson, 2009; Sanborn, Griffiths, & Navarro, 2010). See Gershman and Blei (2012) for a tutorial introduction.

We use a power law kernel $\mathcal{K}(\tau(t) - \tau(t')) = \frac{1}{\tau(t)-\tau(t')}$, with $\mathcal{K}(0) = 0$. This kernel has an important temporal compression property (illustrated in Figure 2). Consider two timepoints, $t_1 < t_2$, separated by a fixed temporal interval, $\tau(t_2) - \tau(t_1)$, and a third time point, $t_3 > t_2$, separated from t_2 by a variable interval, $\tau(t_3) - \tau(t_2)$. In general, because t_3 is closer to t_2 than to t_1, the latent cause that generated t_2 is more likely to have also generated t_3, as compared to the latent cause that generated t_1 having generated t_3 (the contiguity principle). However, this advantage diminishes over time, and asymptotically disappears: as t_1 and t_2 both recede into the past relative to t_3, they become (almost) equally distant from t_3, and it is equally likely that one of their causes also caused t_3.

This completes our description of the animal's internal model. In the next section, we 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.

Associative and Structure Learning

According to our rational analysis, two computational problems confront the animal: (1) associative learning, that is, the adjustment of the model parameters (specifically, the associative weights, $\mathbf{W}$) so that they match the true weights generating outcomes in the environment; a statistically principled way to do this is to adjust the weights so as to maximize the likelihood of the observations seen so far; and (2) structure learning, that is, determining which observation was generated by which latent cause. Here the statistically principled way to determine the assignment of observations to latent causes is to compute the posterior probability for each possible assignment of observations to latent causes. The alternation of these two learning processes can be understood as a variant of the celebrated expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977; Neal & Hinton, 1998). Friston (2005) has argued that the EM algorithm provides a unifying framework for understanding cortical computation.

For our 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 assuming the current weights associated with the different causes are the true weights) and associative learning (the M-step, in which the weights for each cause are updated assuming that the posterior over assignment of observations to latent causes is true):

$$\textbf{E-step}: q_{tk}^{n+1} = P(z_t = k|\mathcal{D}_{1:t}, \mathbf{W}^n) \tag{4}$$

$$\textbf{M-step}: w_{kd}^{n+1} = w_{kd}^n + \eta x_{td}\delta_{tk}^{n+1} \tag{5}$$

for all latent causes k and features d, where n indexes EM iterations, η is a learning rate and

$$\delta_{tk}^{n+1} = q_{tk}^{n+1}(r_t - \textstyle\sum_d w_{kd}x_{td}) \tag{6}$$

is the prediction error at time t for latent cause k. The set of weight vectors for all latent causes at iteration n is denoted by $\mathbf{W}^n$, and the history of cues and rewards is denoted by $\mathcal{D}_{1:t} = \{\mathbf{X}_{1:t}, \mathbf{r}_{1:t}\}$, where $\mathbf{X}_{1:t} = \{\mathbf{x}_1, \ldots, \mathbf{x}_t\}$ and $\mathbf{r}_{1:t} = \{r_1, \ldots, r_t\}$. Note that the updates are performed online in an incremental fashion, so earlier timepoints are not reconsidered.

Associative learning (the M-step of the EM algorithm) in our 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 our generalization the animal can form multiple associations depending on the latent causes it infers (Figure 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:

$$P(z_t = k|\mathcal{D}_{1:t}, \mathbf{W}^n) = \frac{P(\mathcal{D}_{1:t}|z_t = k, \mathbf{W}^n)P(z_t = k)}{\sum_j P(\mathcal{D}_{1:t}|z_t = j, \mathbf{W}^n)P(z_t = j)}. \quad (7)$$

The first term in the numerator is the *likelihood*, encoding the probability of the animal's observations under the hypothetical assignment of the current observation to latent cause z, and the second term is the *prior* probability of this hypothetical assignment (Eq. 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 marginalization over the history of previous latent cause assignments, $\mathbf{z}_{1:t-1}$); we therefore use a simple and effective approximation (see Eq. 13).

Because the E and M steps are coupled, they need to be alternated until convergence (Figure 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, we assume that this happens during intervals between trials, up to some maximum number of iterations (under the assumption that after a finite amount of time the animal will get distracted by something new and cease to "contemplate" its past experience). In our simulations, we take this maximum number to be 3. While the qualitative structure of the theory's predictions does not depend strongly on this maximum number, we found this to produce the best match with empirical data. The explanatory role of multiple iterations comes into play when we discuss the Monfils-Schiller paradigm below.

Prediction

Given the learning model above, when faced with a configuration of CSs on trial t, the optimal prediction of the US is given by its expected value when averaging over the possible latent causes that are currently active:

$$\tilde{r}_t = \mathbb{E}[r_t|\mathbf{x}_t, \mathcal{D}_{1:t-1}] = \sum_{d=1}^{D} x_{td} \sum_k w_{kd} P(z_t = k|\mathbf{x}_t, \mathcal{D}_{1:t-1}, \mathbf{W}^n). \quad (8)$$

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 & Niv, 2010; Kakade & Dayan, 2002). In our simulations, we found that while Eq. 8 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. Indeed, there is some evidence from fear conditioning that freezing to a CS is a nonlinear function of shock intensity (Baldi, Lorenzini, & Bucherelli, 2004). We therefore use a nonlinear sigmoidal transformation of Eq. 8 to model the conditioned response:

$$\text{CR} = 1 - \Phi(\theta; \tilde{r}_t, \lambda), \quad (9)$$

where $\Phi(\cdot; \tilde{r}_t, \lambda)$ is the Gaussian cumulative distribution function with mean $\tilde{r}_t$ and variance λ. One way to understand Eq. 9 is that the animal's conditioned response corresponds to its expectation that the US is greater than some threshold, θ. When $\lambda = \sigma_r^2$, Eq. 9 corresponds precisely to the posterior probability that the US exceeds θ:

$$\text{CR} = P(r_t > \theta | \mathbf{x}_t, \mathcal{D}_{1:t}) = \int_\theta^\infty P(r_t | \mathbf{x}_t, \mathcal{D}_{1:t}) dr_t. \tag{10}$$

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

Understanding Extinction and Recovery

Before modeling specific experimental paradigms, in this section we lay out some general intuitions for how our model deals with extinction and recovery. In previous work (Gershman et al., 2010), we argued that the transition from acquisition to extinction involves a dramatic change in the statistics of the animal's sensory inputs, leading the animal to assign acquisition and extinction to different latent causes. The result of this partitioning is that the acquisition associations are not unlearned during extinction, and hence can be later recovered, as is observed experimentally (Bouton, 2004). Thus, according to our model, the key to persistent reduction in fear is to finesse the animal's sensory statistics such that the posterior favors assigning both acquisition and extinction phases to the same latent cause.

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 our analysis applies to CS prediction errors as well. The prediction error plays two roles in our model: it is an associative learning signal that teaches the animal how to adjust its associative weights, and it is a segmentation signal indicating when a new latent cause is active. When the animal has experienced several CS-US pairs during acquisition, it develops an expectation that is then violated during extinction, producing a prediction error. Typical learning rules such as Rescorla-Wagner's are "error-correcting" rules that modify associations or values so as to reduce future prediction errors. In our model, however, the 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 persistent accumulation of these prediction errors over the course of extinction eventually makes the posterior probability of a new cause high. Thus, standard acquisition and extinction procedures eventually lead to the formation of two memories or associations, one for CS-US and one for CS-noUS.

The opposing effects of prediction errors on associative and structure learning are illustrated in Figure 3. If the prediction errors are too small, the posterior probability of the acquisition

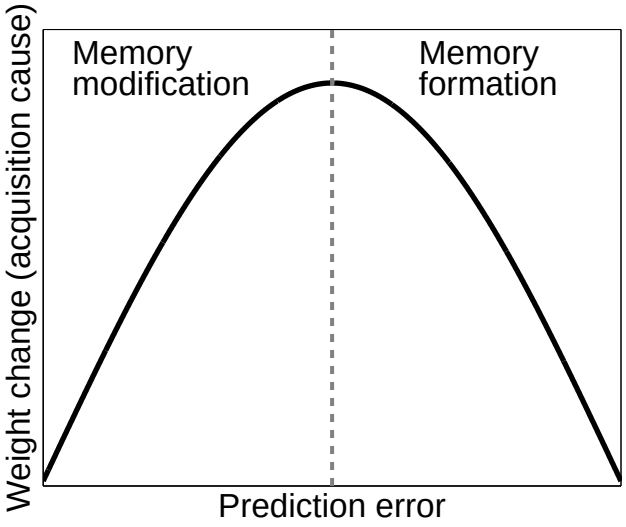

Figure 3: Cartoon of the model's predictions for fear extinction. The X-axis represents the size of the prediction error (experienced minus expected US) during extinction, and the Y-axis represents the change (after learning) in the weight corresponding to the "acquisition latent cause" (i.e., the latent cause inferred by the animal during conditioning).

latent cause will be high (leading to modification of the original memory) but the amount of CS-US weight change will be small as there is little discrepancy between what was predicted and what was observed; if the prediction errors are too big, the posterior probability of the acquisition latent cause will be low (leading to formation of a new memory), and the change in the weight corresponding to the original memory will again be small. In theory, therefore, there should exist 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. Later we describe a behavioral paradigm (the Monfils-Schiller paradigm) that achieves this sweet spot.

To get a feeling for how the model's response to prediction errors depends on the parameter settings, consider a simple conditioning paradigm in which a single cue has been paired N times with reward ($\mathcal{D}_{1:N} = \{x_t = 1, r_t = 1\}_{t=1}^{N}$), with a constant ITI ($\tau(t) - \tau(t-1) = 1$). Under most parameter settings, this will result in all the acquisition trials being assigned to a single latent cause (hence we ignore the cause subscript k in this example and refer to the single cause as the "acquisition latent cause"). Now consider what happens when a single extinction trial ($x_{N+1} = 1, r_{N+1} = 0$) is presented. The posterior over latent causes (Eq. 4) is proportional to the product of 3 terms: (1) The prior over latent causes, (2) the likelihood of the US, and (3) the likelihood of the CS. The third term plays a negligible role, since the CS is constant across acquisition and extinction, and hence no CS prediction error is generated. As N grows, the prior probability of the acquisition latent cause increases, due to the simplicity bias of the Chinese restaurant process. However, associative learning

of the weight vector counteracts this effect, since the US expectation, and hence the size of the prediction error due to the absence of the US (encoded in the likelihood term), also grows with N, asymptoting once the US prediction is fully learned. This can produce, under certain parameter settings, a posterior over latent causes that changes non-monotonically with N; sensitivity to the US prediction error increases as σ_r^2 decreases (higher confidence in US predictions) and α increases (weaker simplicity bias).

In order to understand some of the empirical phenomena described below, we must also explain why spontaneous recovery occurs in our model: Why does the posterior probability of the acquisition cause increase as the extinction-test interval is lengthened? The answer lies in our choice of temporal kernel $\mathcal{K}$ 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 acquisition and test increases.[2] Thus, the advantage of the extinction cause over the acquisition cause at test diminishes with the extinction-test interval. One implication of this analysis is that spontaneous recovery should never be complete, since the prior probability of the acquisition cause can never *exceed* the probability of the extinction cause (though the ratio of probabilities increases monotonically towards 1 as the extinction-test interval increases); this appears generally consistent with empirical data (Rescorla, 2004).

Boundary Conditions on Reconsolidation

In this section, we explore several boundary conditions on reconsolidation (see Nader & Hardt, 2009, for a review). Our goal is to show that these conditions fall naturally out of our rational analysis of Pavlovian conditioning. We seek to capture the *qualitative* pattern of results, rather than their precise quantitative form. We thus use the same parameter settings for all simulations (see Appendix), rather than fitting the parameters to data for each particular case. The parameter settings were chosen heuristically, but our results hold over a range of values.

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

[2] A similar idea was used by Brown, Neath, and Chater (2007) in their model of episodic memory to explain recency effects in human list learning experiments.

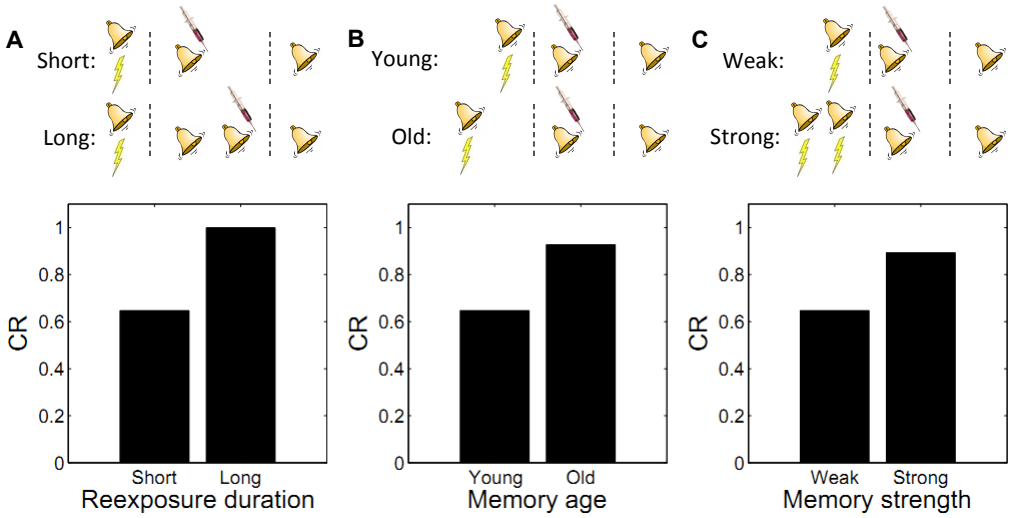

Figure 4: **Boundary conditions on reconsolidation**. Top row shows a schematic of the experimental design (bell represents the tone CS, lightning bolt represents the shock US, syringe represents the injection of a protein synthesis inhibitor), with a conditioning → extinction → test structure. Bottom row shows model predictions in the test phase. Memory updating is attenuated under conditions of (A) longer reexposure, (B) older or (C) stronger memories.

The trace dominance principle

Using fear conditioning in the Medaka fish, Eisenberg et al. (2003) found that applying an amnestic agent after a single reexposure to the CS (i.e., a single extinction trial) caused retrograde amnesia for the reactivated fear memory. In contrast, applying the amnestic agent after multiple reexposures caused retrograde amnesia for extinction, that is, high recovery of fear was observed in a test on the following day. Similar results have been obtained with mice (Suzuki et al., 2004), rats (J. Lee, Milton, & Everitt, 2006), and the crab *Chasmagnathus* (Pedreira & Maldonado, 2003). A trace-dominance principle interpretation of these data suggests that reexposure duration determines the dominance of a memory. This is also seen in our model: short reexposure duration (operationalized by a single CS presentation) favors assignment of the reexposure trial to the acquisition latent cause. This follows from the simplicity bias in the latent cause prior: In the absence of strong evidence to the contrary, new observations are preferentially assigned to previously inferred causes. However, with longer durations of extinction (e.g., two or more CS presentations), evidence favoring a new latent cause (accruing from persistent prediction errors) increases, favoring assignment of these trials to a new latent cause (the 'extinction' cause). This logic leads to model predictions consistent with the empirical data (Figure 4A).

Memory age

By manipulating the interval between acquisition and reexposure, Suzuki et al. (2004) demonstrated that the amnestic effects of PSI injection were more pronounced for young memories (i.e., short intervals). Winters, Tucci, and DaCosta-Furtado (2009) found a similar effect with the NMDA receptor antagonist MK-801 administered prior to reexposure, 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 our model: old observations are less likely (under the prior) to have been generated by the same latent cause as recent observations. Thus, there is an inductive bias against modifying old memory traces. Figure 4B shows simulations of the Suzuki paradigm, demonstrating that our 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 S. Wang, de Oliveira Alvares, & Nader, 2009). Specifically, increasing the number of acquisition trials led to persistent fear even after reexposure to the CS and PSI injection. In terms of our 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. Because the weight is large, the prediction error is large, which causes the model to infer a new cause for the CS-alone trial. This new cause, in turn, would be the cause undergoing weakening due to PSI administration (i.e., the trace-dominance principle), rather than the old cause associated with the fear memory. Simulations of this experiment (Figure 4C) demonstrate that stronger memories are more resistant to updating in our 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 of the fear memory (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 more 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 our theory, due to the time-dependent prior over latent causes, which prefers assigning trials separated by a long temporal interval to different causes. As a result, 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 "acquisition" latent cause as compared to a yet another new latent cause, different from that

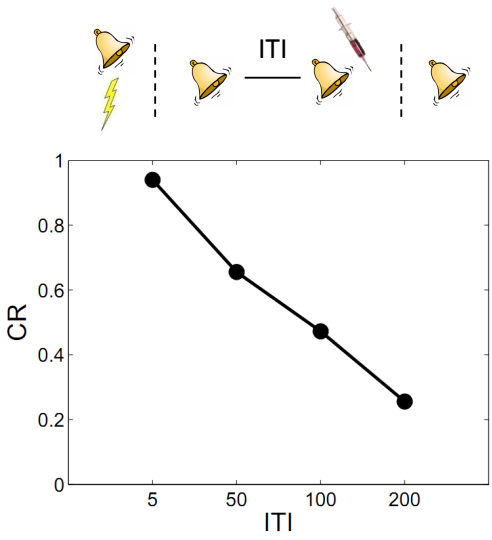

Figure 5: **Timing of multiple reexposures**. Lengthening the intertrial interval (ITI) between multiple reexposures increases the effectiveness of PSI administration in attenuating fear at test. Depicted are simulation results.

generating the first reexposure trial and from the cause that generated the acquisition trials (Figure 5).[3]

Prediction error and novelty

As described above, prediction errors play two central roles in our model, driving both associative and structure learning. Of particular relevance to this claim is research showing that violation of the animal's expectations (i.e., prediction error) is necessary to induce memory updating (Morris et al., 2006; Pedreira, Pérez-Cuesta, & Maldonado, 2004; Sevenster, Beckers, & Kindt, 2013; Winters et al., 2009). In one manifestation of this boundary condition, Pedreira et al. (2004) found that a PSI administered after a reinforced retrieval trial did not diminish the original fear memory, suggesting that in this case reconsolidation of the retrieved memory was relatively intact despite the PSI. This finding is consistent with our model (Figure 6A), which predicts that the acquisition cause will be retrieved, but the PSI-induced decrement will be offset by the reinforcement. 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 acquisition if the object is paired with novelty. We simulated this experiment by adding a novel CS during the retrieval

[3]This prediction is to some extent parameter-dependent: If the concentration parameter is sufficiently large, then a new latent cause will be favored.

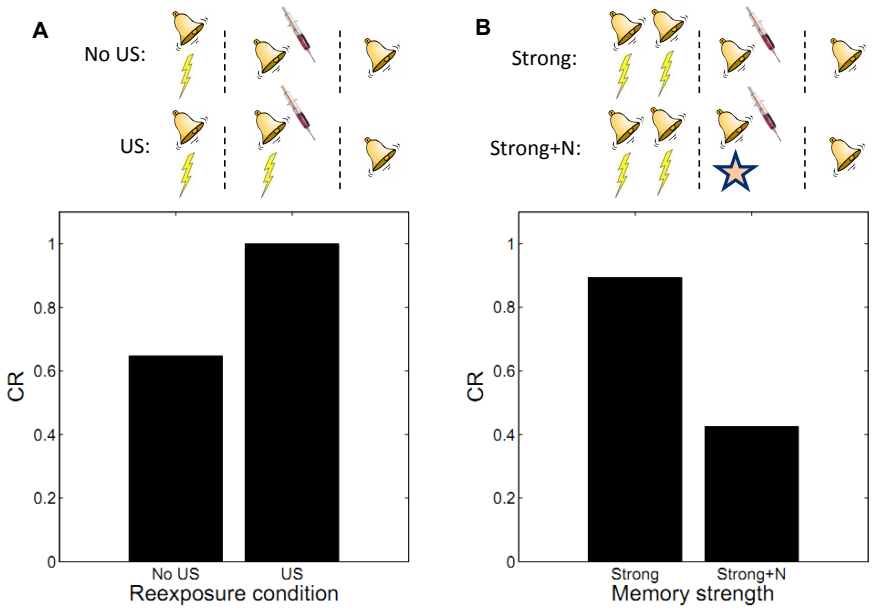

Figure 6: **The role of prediction error in reconsolidation**. (A) Presenting the unconditioned stimulus (US) during reexposure offsets the PSI-induced decrement of the acquisition cause. (B) Strong memories can be updated when reexposure is accompanied by a novel object (Strong+N, indicated by a star), thereby eliminating the strength-based boundary condition.

trial; Figure 6B shows that a strong memory is sensitive to disruption when accompanied by novelty.

Cue-specificity

Doyère, Debiec, Monfils, Schafe, and LeDoux (2007) reported that disruption of reconsolidation 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 7A). This finding is explained by our model by observing that learning only affects the CSs associated with the current inferred latent cause. In a recent study, Debiec, Diaz-Mataix, Bush, Doyère, and LeDoux (2013) showed that cue-specificity of reconsolidation depends on separately training the two CSs; when they are trained in compound, reactivating one CS can render the other CS labile. Our model reproduces this effect (Figure 7B) as in this case the compound cue is assigned to a single latent cause, thereby coupling the two CSs.

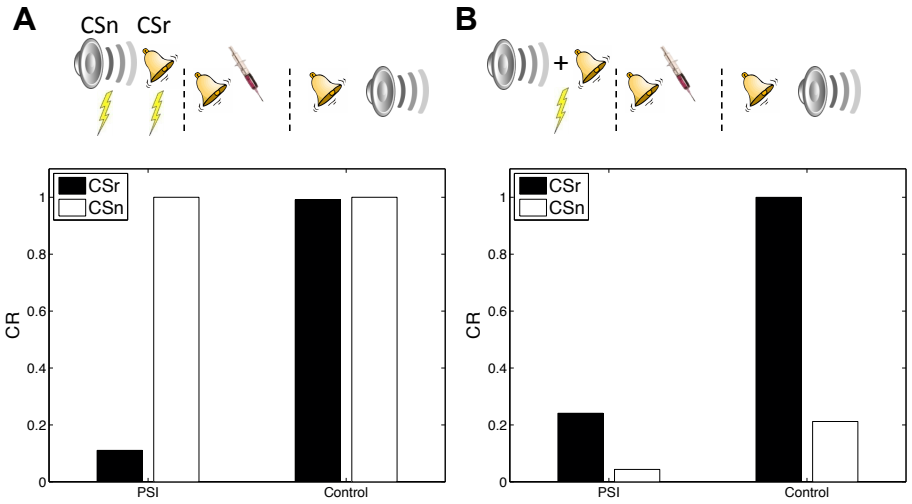

Figure 7: **Cue-specificity of amnestic treatment**. (*A*) Disruption of reconsolidation by amnestic treatment affects the reactivated cue (CSr) but not the non-reactivated cue (CSn). (*B*) When trained in compound, reactivating CSr renders CSn vulnerable to disruption of reconsolidation.

Transience of amnesia

A major focus of theories of experimental amnesia (i.e., forgetting of the association formed during acquisition) has been the observation that, under a variety of circumstances, recovery from amnesia can occur (Miller & Matzel, 2006; Riccio et al., 2006). A study by Power, Berlau, McGaugh, and Steward (2006) provides a clear demonstration: Following conditioning, post-retrieval intrahippocampal infusions of the PSI anisomycin reduced conditioned responding when the rats were tested 1 day later, but responding recovered when the test was administered after 6 days. Thus, the PSI-induced memory impairment was transient (see also Lattal & 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 diminished, but did not entirely eliminate, the reactivated memory, then subsequent recovery could reflect new learning added on to the residual memory trace.[4]

The explanation that our theory offers for this phenomenon is related to Gold's interpretation, in that it also assumes a residual memory trace. Since the amnestic agent does not entirely eliminate the memory trace, later recovery of the fear memory occurs because the relative probability of assigning a new test observation to the acquisition cause rather than to the cause associated with the retrieval session (which was, in effect, a short extinction session), increases over time (a consequence of temporal compression by the power law kernel, as explained above). In other words, this is a form of spontaneous recovery: The original

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

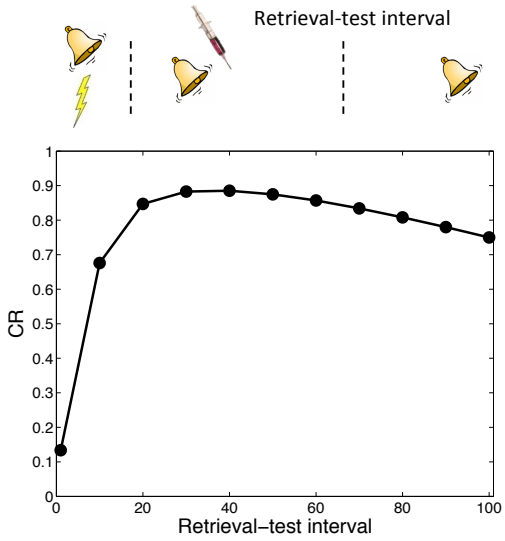

Figure 8: **Transience of amnesia**. Lengthening the interval between retrieval and test produces recovery from amnesia.

(weak) memory becomes more retrievable over time. Simulations shown in Figure 8 demonstrate that this explanation can account for the increase in CR with longer retrieval-test intervals. Interestingly, the model predicts that further increasing the retrieval-test interval will eventually result in reduced responding, because of the increased probability of a new latent cause at test.

The Monfils-Schiller Paradigm

Two influential studies (Monfils et al., 2009; Schiller et al., 2010) demonstrated that a single reexposure ("retrieval trial") of a CS that had been associated with a shock, 10-60 minutes before extinction, leads to persistent reduction of fear as measured by renewal, reinstatement and spontaneous recovery tests. Importantly, this effect did not require pharmacological or other interventions such as PSIs, and it was evident in both rodents (Monfils et al., 2009) and humans (Schiller et al., 2010). These studies also revealed that: (1) reduction of fear lasts up to a year later; (2) the reduction was specific to the cue-reactivated memory; and (3) increasing the retrieval-extinction interval to 6 hours eliminates the effect. That is, extinction after a retrieval trial is more effective at modifying the original association than regular extinction, but this only holds for extinction sessions administered relatively promptly after the retrieval cue. This latter finding suggests that the retrieval cue engages a time-limited plasticity window, in which extinction operates. These findings have been replicated several times in rats (Baker, McNally, & Richardson, 2013; Clem & Huganir, 2010; Jones, Ringuet, & Monfils, 2013; Olshavsky, Jones, Lee, & Monfils, 2013; Olshavsky, Song, et al., 2013)

and humans (Agren et al., 2012; Oyarzún et al., 2012; Schiller, Kanen, LeDoux, Monfils, & Phelps, 2013; Steinfurth et al., 2014), though the generality of the paradigm remains controversial (W. Chan, Leung, Westbrook, & McNally, 2010; Costanzi, Cannas, Saraulli, Rossi-Arnaud, & Cestari, 2011; Kindt & Soeter, 2013; Soeter & Kindt, 2011). We address some of the inconsistencies across studies below.

It is important to recognize that the so-called "retrieval trial" is operationally no different from an extinction trial—it is a CS presented alone. Essentially, the principal salient difference between the Monfils-Schiller paradigm and regular extinction training is that in the Monfils-Schiller paradigm, the interval between the 1st and 2nd extinction trials is substantially longer than the intervals between all the other trials. Another difference (which we do not explicitly address here) is that in the Monfils-Schiller paradigm, the subject spends the retrieval-extinction interval outside the acquisition context, in its home cage. This phenomenon is thus puzzling for most—if not all—theories of associative learning. What happens during this one interval that dramatically alters later fear memory?

Model simulations of the Monfils-Schiller paradigm are shown in Figure 9. We simulated 3 conditions, differing only in the retrieval-extinction interval (REI): *No Ret* (REI=0), *Ret-short* (REI=3), *Ret-long* (REI=100).[5] As observed experimentally, in our simulations 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, in the Ret-short condition there was no spontaneous recovery of fear at test. Examining the posterior distributions over latent causes in the different conditions (Figure 9B-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 acquisition cause in the Ret-short condition.

Our theoretical explanation of data from the Monfils-Schiller paradigm rests critically on what happens during the interval between the 1st and 2nd extinction trials, that is, the REI. Since there is some probability that the original ("acquisition") latent cause is active during the REI, the first iteration of associative learning in the EM algorithm will reduce the CS-US association for that latent cause, due to the assumption that that cause also generated the CS-alone "retrieval" trial. This, in turn, makes it incrementally more likely that the original latent cause is active (since the prediction error decreases with each reduction of the associative strength), so on the next EM iteration the CS-US association is further reduced. In our model, this process is repeated for a maximum of 3 iterations (see the model description for a justification of this cutoff), and the number of iterations depends on the length of the REI. 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 acquisition cause. In contrast, in the Ret-short condition, the probability that the retrieval trial is assigned to the acquisition latent cause increases over the course of the interval. The ensuing extinction trials are then more likely to be assigned to the same latent cause, therefore decreasing the CS-US association even more. Spontaneous recovery of the original

[5]Time is measured in arbitrary units here; see the Appendix for a description of how these units roughly map on to real time.

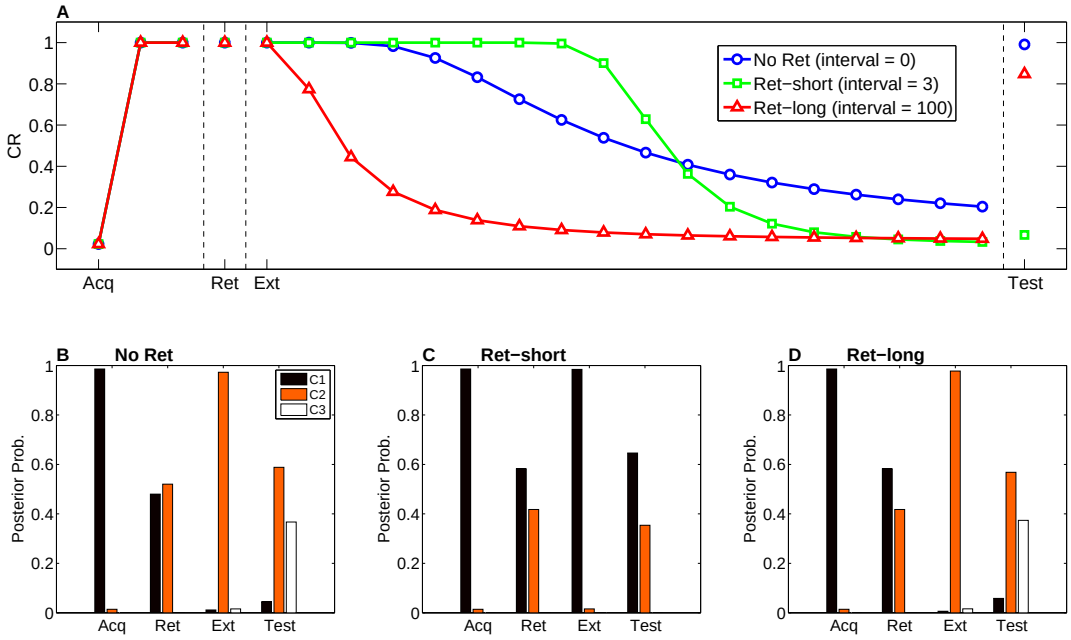

Figure 9: **Model predictions for the Monfils-Schiller paradigm**. (A) Simulated conditioned response (CR) during acquisition (Acq; 3 CS-US pairs), retrieval (Ret; 1 CS presentation 24 hours after acquisition, followed by no interval, a short interval, or a long interval before the next phase), extinction (Ext; CS-alone presentations) and a test phase 24 hours later. Three conditions are shown: No-Ret (no interval between retrieval and extinction; the "Ret" trial depicted here is the first trial of extinction), 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. Probabilities for only the top 3 highest-probability causes are shown here.

fear memory at test is thus attenuated due to modification or "erasure" of the acquisition latent cause's CS-US association (Figure 10A). Lastly, when the retrieval-test interval is long (as in the Ret-long condition), although the CS-US association of the acquisition latent cause is reduced during the REI, extinction trials will be preferentially assigned to a new latent cause due to the time-sensitive prior that suggests that events far away in time are generated by different causes. Thus the original association is not significantly attenuated, and spontaneous recovery of fear occurs at test. This nonmonotonic dependence on the retrieval-test interval is shown quantitatively in Figure 10B.

Figure 11 shows simulations of the cue-specificity experiment reported in Schiller et al. (2010). In a within-subjects design, two CSs were paired with shock, but only one was reexposed prior to extinction in a "retrieval" trial. Consistent with the findings 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 our 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-extinction interval suggests that distracting or occupying animals during the interval should disrupt the Monfils-Schiller effect. For example, our theory predicts that giving rats a secondary task to perform during the interval will prevent the iterative weakening of the CS-US association of the acquisition cause, leading to assignment of extinction trials to a new latent cause (as in regular extinction) and 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.

Boundary Conditions in the Monfils-Schiller Paradigm

Several studies have reported recovery of memory in the Monfils-Schiller paradigm (W. Chan et al., 2010; Costanzi et al., 2011; Kindt & Soeter, 2013; Soeter & Kindt, 2011). Is the paradigm inherently fragile, or do these discrepancies delineate systematic boundary conditions? Auber, Tedesco, Jones, Monfils, and Chiamulera (2013) identified many methodological differences between experiments using the Monfils-Schiller paradigm. The question facing our theory is whether the effects of these differences can be explained as a rational consequence of inference given sensory data. Many of the differences involve experimental variables that are outside the scope of our theory, such as the tone frequency in auditory fear conditioning (W. Chan et al., 2010), or affective properties of picture stimuli in human studies (Kindt & Soeter, 2013; Soeter & Kindt, 2011). We therefore focus in this section on two methodological differences that fall within the scope of our theory.

Costanzi et al. (2011) trained mice to associate a foot-shock with a context, and then induced retrieval of the contextual memory 29 days later by placing the mice in the conditioning context for 3 minutes. An extinction session in the same context followed one hour later.

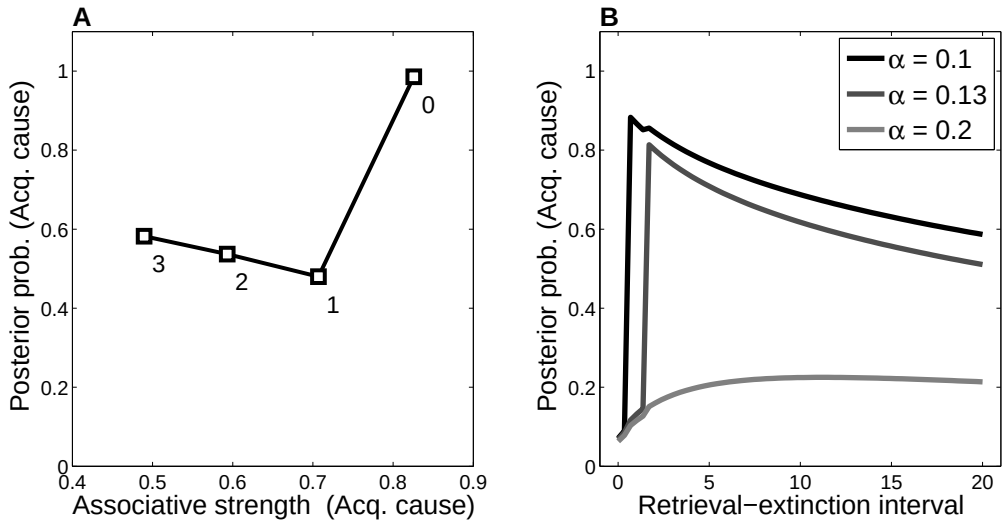

Figure 10: **Dynamics of associative and structure learning during the retrieval-extinction interval**. (A) The X-axis represents the associative weight corresponding to the acquisition latent cause. The Y-axis represents the posterior probability that the acquisition latent cause is active for the retrieval trial. Each numbered square indicates a particular iteration during the retrieval-extinction interval; the square numbered "0" indicates the last trial of acquisition. Initially, the prediction error causes the posterior to favor a new latent cause rather than the old acquisition cause, however, 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 acquisition cause. (B) As the retrieval-extinction interval grows longer, the probability of assigning the first extinction trial to the acquisition cause changes non-monotonically. Two non-reinforced trials very close in time are likely to come from a new latent cause, thus the posterior probability of the acquisition cause generating these trials starts low. It peaks at a larger retrieval-extinction interval; as this interval increases, the acquisition cause's associative strength is incrementally reduced, thereby making the extinction trials more likely under the acquisition cause. The curve then gradually diminishes due to the time-sensitive prior that causes temporally separated events to be more likely to be generated by different causes (Eq. 3). The specific form of this curve is strongly influenced by the probability of inferring a new latent cause, that is, by the value of α; curves here illustrate results for three different values of α.

The next day, the mice were tested for contextual fear in the conditioning context. Costanzi et al. (2011) found that extinction after retrieval did not attenuate contextual fear, contrary to the findings of Monfils et al. (2009).

Auber et al. (2013) pointed out that a crucial difference between the studies of Costanzi et al. (2011) and Monfils et al. (2009) was the acquisition-retrieval interval: 29 days in Costanzi

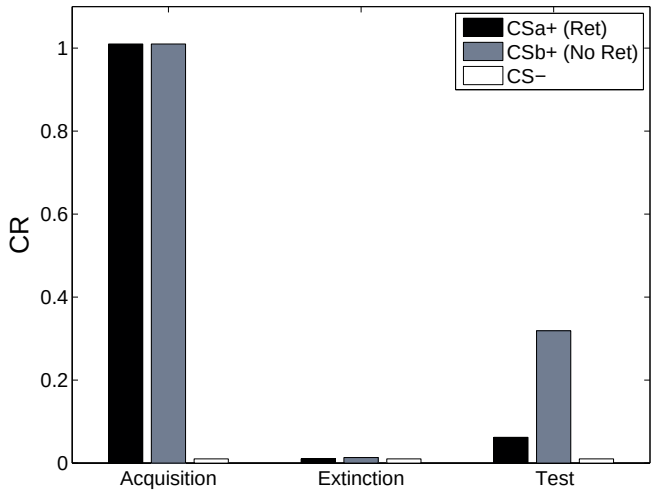

Figure 11: **Cue-specificity in the Monfils-Schiller paradigm**. Model simulations of the within-subjects design reported by Schiller et al. (2010), in which two CSs (CSa+ and CSb+) were individually paired with shock (CS− was never paired with a shock), but only one (CSa+) was reexposed in a "retrieval" trial prior to extinction. Fear recovery is attenuated for the reexposed CS.

et al. (2011) and 1 day in Monfils et al. (2009). As we reviewed above, it is well-established that older memories resist modification (Alberini, 2007). According to our theory, this phenomenon occurs because when the acquisition-retrieval interval is long, the retrieval trial is less likely to have been generated by the same latent cause as the acquisition trials. Our theory predicts the same effect in the Monfils-Schiller paradigm (Figure 12A), consistent with the findings of Costanzi et al. (2011).

In another study reporting discrepant results, W. Chan et al. (2010) found that the Monfils-Schiller paradigm failed to prevent the return of fear in renewal and reinstatement tests. Auber et al. (2013) observed that the study by W. Chan et al. (2010) used different experimental boxes located in different rooms for acquisition and for retrieval and extinction, whereas in their renewal experiment Monfils et al. (2009) modified the conditioning box to create a new context for retrieval and extinction. We simulated the different contexts by adding a "context" feature that allowed us to parametrically vary the similarity between acquisition and retrieval/extinction contexts. In particular, we assumed that this feature was 1 in acquisition, and then in retrieval and extinction we represented the similar context by setting the feature to 0.8, whereas the dissimilar context feature was set to 0. We found that retrieval and extinction in a similar context led to less renewal at test than did retrieval and extinction in a very different context (Figure 12B). This is because when acquisition and retrieval/extinction contexts are similar, there is a higher probability that the latter

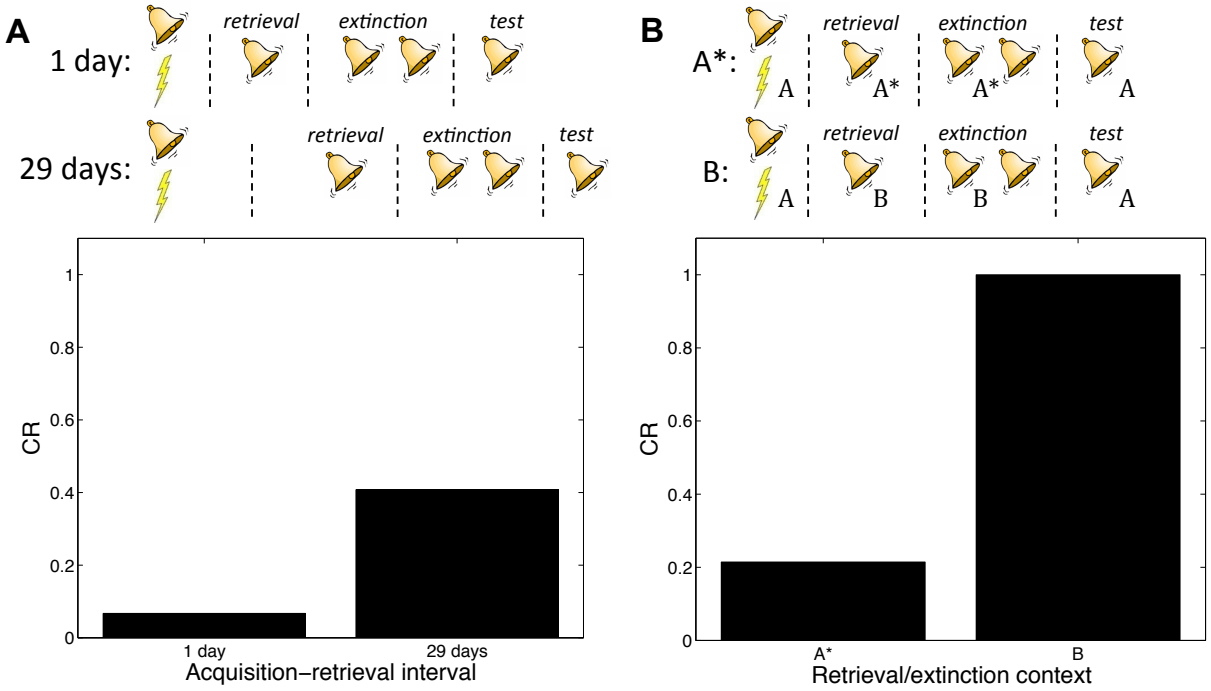

Figure 12: **Boundary conditions in the Monfils-Schiller paradigm**. (*A*) A short acquisition-retrieval interval is more effective at attenuating spontaneous recovery of fear than a long acquisitio-retrieval interval. (*B*) A retrieval/extinction context (A*) that is similar to the acquisition context (A) leads to attenuated renewal of fear when tested in A, whereas a very dissimilar context (B) leads to renewal.

trials will be assigned to the original acquisition latent cause—i.e., that the "retrieval" trial will indeed retrieve the old association (see also Gershman et al., 2010; Gershman, Schapiro, Hupbach, & Norman, 2013).

In summary, our theory reconciles some of the discrepant findings across studies using the Monfils-Schiller paradigm. These findings can therefore be regarded as delineating systematic boundary conditions on the original findings, although more work will be required to ascertain whether all the discrepancies identified by Auber et al. (2013) can be rationalized in this way.

Testing a Prediction of the Model: Extinction Prior to Retrieval Attenuates the Monfils-Schiller Effect

Our model predicts that subjects in the Monfils-Schiller paradigm retrieve the latent cause responsible for the acquisition trials and update it 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. As a result, extinction training will not modify the original memory trace (i.e., weaken the originally acquired CS-US association) because the acquisition cause will not be 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 investigate this prediction experimentally, we 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 our model, we hypothesized that the E-R-E group would infer a new latent cause for unreinforced trials, while the R-E group would not, and that this would interact with the ability of a future retrieval+extinction session to modify the memory of the original acquisition cause.

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.

Apparatus and Stimuli

All behavioral procedures took place in a single context—a standard conditioning chamber 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 (ITIs) = 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, 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).

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.

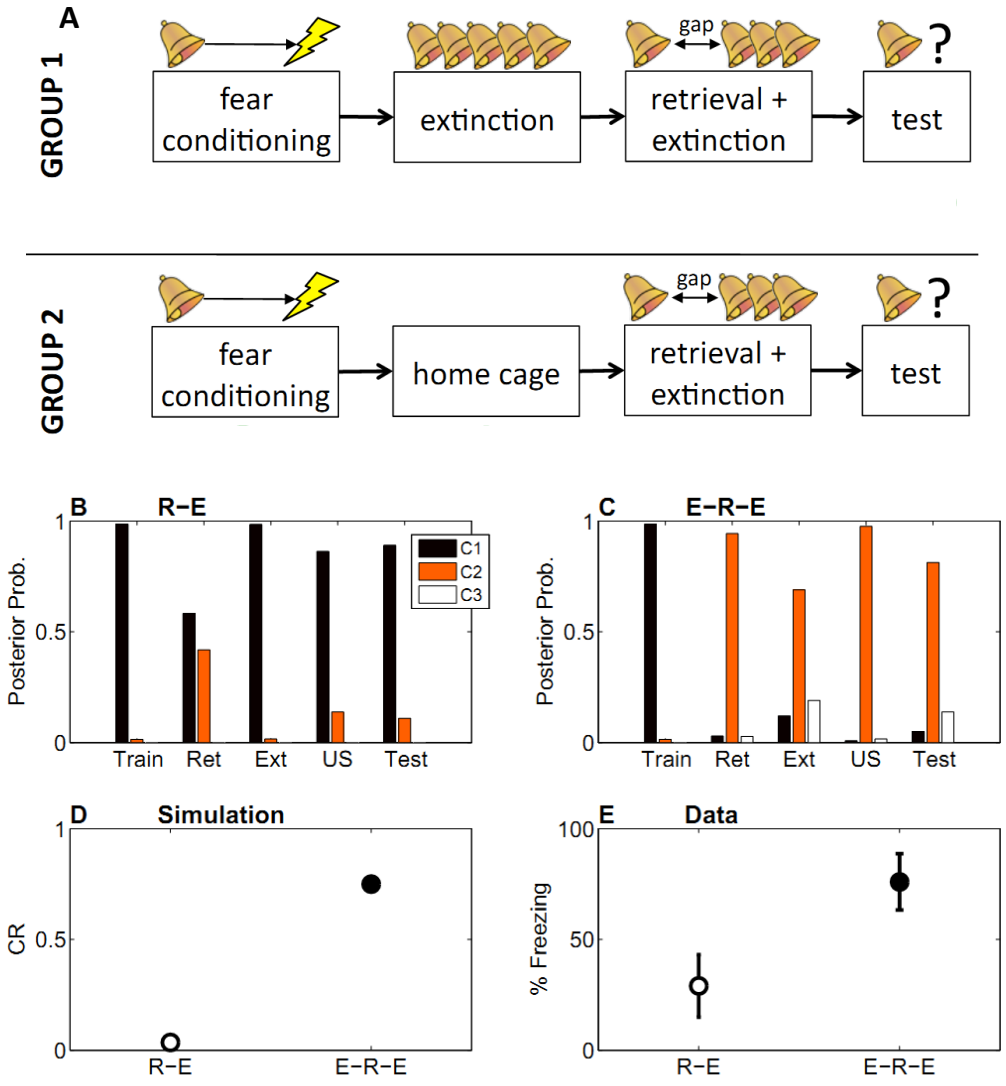

Figure 13: **Performing extinction prior to retrieval attenuates reconsolidation**. (A) Experimental design: Rats were fear-conditioned rats using 3 tone-shock pairings. On the next day, one group of rats (E-R-E) received an extinction session (5 non-reinforced tone CSs) while a second group (R-E) did 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 2 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.

Results

Our experimental results (Figure 13E) were in accord with the model's predictions: rats that received an extinction session prior to the retrieval+extinction session (the E-R-E group) showed greater reinstatement compared to the R-E group that did not receive this initial extinction session [$t(9) = 2.29, p < 0.05$]. This constitutes a new boundary condition on reconsolidation: *the original fear memory is not updated when the retrieval trial is preceded by extinction*. Our 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.

It should be noted that our results appear to contradict a recent study by Millan, Milligan-Saville, and McNally (2013), which found that performing retrieval after extinction results in reduced reinstatement. Millan et al. (2013) suggested that, in this case, rather than rendering the acquisition memory labile, the retrieval manipulation facilitates discrimination between acquisition and extinction memories (cf. Capaldi, 1994; Redish et al., 2007). There are many procedural differences between our experiment and those of Millan et al. (2013); for example, Millan et al. (2013) used an alcohol seeking paradigm instead of fear conditioning, and retrieval was performed after the entire extinction phase was completed. More research is needed to determine whether their results extend to fear conditioning.

Discussion

We have shown how major phenomena in 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: a *structure learning* process that infers the latent causes underlying sensory inputs, and an associative learning process that adjusts the parameters of the internal model. Our theory expands on previous associative learning theories that concentrate on the latter process while assuming a fixed model of the environment, and formalizes the dynamic interplay between memory processes and learning processes, which is at the heart of reconsolidation phenomena. In particular, we showed that the interplay between structure learning and associative learning can reproduce parametric variations in the effects of reconsolidation treatments, consistent with experimentally observed boundary conditions.

Explaining the Mystery of the Monfils-Schiller Paradigm

One of the most intriguing findings in the reconsolidation literature in recent years was the discovery of a noninvasive behavioral treatment that is effective at attenuating recovery of conditioned fear (Agren et al., 2012; Monfils et al., 2009; Schiller et al., 2010). Monfils,

Schiller and their colleagues demonstrated (in both rats and humans) that performing extinction acquisition within a short interval following a retrieval cue (an unreinforced CS presentation) reduced later recovery of fear. The effect was later demonstrated in appetitive learning (Ma, Zhang, & Yu, 2012) and contextual fear conditioning (Flavell, Barber, & Lee, 2011; Rao-Ruiz et al., 2011). The Monfils-Schiller paradigm has also be applied to drug-associated memory, attenuating drug-seeking in rats and cue-induced heroin craving in human addicts (Xue et al., 2012), as well as reducing cocaine-primed reinstatement of conditioned place preference (Sartor & Aston-Jones, 2013) and context-induced reinstatement of alcoholic beer seeking (Millan et al., 2013) in rats.

The Monfils-Schiller paradigm is theoretically tantalizing because it is not *a priori* clear what is the difference between the retrieval trial and the first trial of any extinction session—why is it that the CS-alone trial in the Monfils-Schiller paradigm acts as a "retrieval cue" that initiates a reconsolidation process, while the first CS-alone trial of a regular extinction session does not? The effectiveness of this paradigm thus seems to challenge our basic understanding of the interplay between learning and memory processes. Our model explains this puzzle by stressing the role of the extended period of learning (in our model, additional iterations of the EM algorithm) during the long retrieval-extinction gap, in which the rat is left in its home cage to essentially "ruminate" about its recent experience. Thus our explanation rests not on the existence of a separate reconsolidation process that is invoked by the retrieval trial, but rather on the same learning and memory mechanisms that are at play in acquisition and in extinction—the idea that inference about the latent structure of the environment affects whether new information will update an old association, or whether it will be attributed to a new memory (new latent cause). In this sense, according to our theory the "retrieval" trial is, in fact, not different from any other trial, and perhaps a more accurate nomenclature would be to call the post-first-CS gap an "updating gap".

Despite its successes, the effectiveness of the Monfils-Schiller paradigm has been controversial, with several replication failures (W. Chan et al., 2010; Costanzi et al., 2011; Ishii et al., 2015; Kindt & Soeter, 2013; Ma et al., 2012; Soeter & Kindt, 2011). Auber et al. (2013) described a number of subtle methodological differences between these studies, possibly delineating boundary conditions on the Monfils-Schiller paradigm. Inspired by this suggestion, we showed through simulations that the effects of several methodological differences (acquisition-retrieval interval and context similarity) are indeed predicted by our theory. Nevertheless, important boundary conditions on the length and characteristics of the updating gap remain to be studied; for instance, does it have to be longer than 10 minutes (as has been done in previous experiments) or is the minimum length of this gap more parametrically dependent on the overall pace of new information (e.g., the length of the ITIs at acquisition).

From a neurobiological standpoint, recent work has lent plausibility to the claim that the Monfils-Schiller paradigm erases the CS-US association learned during acquisition. 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, LeDoux, Zador, & Malinow, 2005), suggesting that changes in post-synaptic AMPA receptor density may be the neural substrate of associative learning in fear conditioning. Monfils et al. (2009) reported increased phosphorylation of AMPA receptors in the lateral amygdala after the retrieval trial (a possible correlate of memory labilization), and Clem and Huganir (2010) found that extinction following retrieval resulted in synaptic removal of calcium-permeable AMPA receptors. The latter finding is significant in that it indicates a reversal of the synaptic changes that occurred during conditioning, supporting the view that the Monfils-Schiller paradigm results in unlearning of the original CS-US association.

Our theoretical analysis is consistent with these findings. We 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. Evidence suggests that associative learning in fear conditioning relies on the basolateral nucleus of the amygdala (Blair, Schafe, Bauer, Rodrigues, & LeDoux, 2001), whereas structure learning may be supported by the hippocampus (Aggleton, Sanderson, & Pearce, 2007; Gershman et al., 2010; Gershman, Radulescu, Norman, & Niv, 2014). Thus, deactivating the amygdala or stimulating the hippocampus (e.g., using optogenetic manipulations) should block memory updating.

In a behavioral experiment, we examined another prediction of our 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. Our behavioral data confirmed 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, LeDoux, & Nader, 2002), while others have not (Biedenkapp & 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, Bell, Ramirez, Eaddy, & Su, 2009; Milton, Lee, Butler, Gardner, & Everitt, 2008), while others have not (Hernandez & 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.

A Neural Circuit for Reconsolidation

Although we have so far not committed to any specific neural implementation of our model, we believe it fits comfortably into the computational functions of the circuit underlying Pavlovian conditioning. We propose a provisional mapping onto this circuit, centering on the amygdala and the "hippocampal-VTA loop" (Lisman & Grace, 2005) connecting the hippocampus and the ventral tegmental area in the midbrain. Our 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, we have suggested that the hippocampus is a key brain region involved in partitioning the world into latent causes (Gershman et al., 2010, 2014). This view resonates with earlier models emphasizing the role of the hippocampus in encoding sensory inputs into a statistically compressed latent representation (Fuhs & Touretzky, 2007; Gluck & Myers, 1993; Levy, Hocking, & Wu, 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 & Good, 1993), indicating that animals without a hippocampus cannot "carve nature at its joints" (i.e., partition observations into latent causes; see Gershman & Niv, 2010; Gershman et al., 2015).

Within the current model, we propose that the dentate gyrus (DG) activates latent 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. 3). Mechanistically, these representations may be encoded in attractors by the dense recurrent collaterals that are characteristic of CA3 (McNaughton & Morris, 1987).

An important aspect of our 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, Wiles, and Gage (2006) have suggested that immature granule cells, by virtue of their low activation thresholds, high resting potentials and constant turnover, cause inputs that are distant in time to map onto distinct CA3 representations. Furthermore, evidence suggests that new granule cells die over time if they are not involved in new learning (Shors, Anderson, Curlik, & Nokia, 2012), offering another mechanism by which the contiguity principle could be implemented.

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, Johansen, & Blair, 2011). Accordingly, we 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 review), and has been specifically implicated in Pavlovian fear conditioning (Pezze & Feldon, 2004), although little is known about the phasic firing properties of dopamine neurons during aversive conditioning. Dopamine gates synaptic plasticity in the BLA (Bissière, Humeau, & Luthi, 2003), consistent with its hypothesized role in driving the learning of CS-US associations. We 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 & Grace, 2005), potentially providing a mechanism by which the posterior distribution over latent causes can modulate the prediction errors, as suggested by our model. In appetitive conditioning experiments, Reichelt, Exton-McGuinness, and Lee (2013) have shown that dysregulating dopaminergic activity in the VTA prevented the destabilization of memory by NMDA receptor antagonists (injected systemically following a retrieval trial), consistent with the hypothesis that dopaminergic prediction errors are necessary for memory updating after memory retrieval. It is not known whether this effect is mediated by dopaminergic projections to the hippocampus.

Comparison to a Mismatch-based Autoassociative Neural Network

Osan, Tort, and Amaral (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, Osan, Roesler, & Tort, 2008, for a related model). In this model, acquisition and extinction memories correspond to attractors in the network, formed through Hebbian learning on the retrieved attractor. 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 (i.e., the weights are adjusted to favor the input pattern). Mismatch is assumed to accumulate over the course of the input presentation.

The degradation process can be viewed as a kind of error-driven learning—when the network does not accurately encode the current input, the weights are adjusted to encode it more accurately in the future. In the case of extinction, this implements a form of unlearning. The relative balance of Hebbian learning and mismatch-induced degradation determines the outcome of extinction training. Assuming that the original shock pattern is retrieved at the beginning of extinction, degradation weakens the shock pattern, whereas Hebbian learning strengthens the retrieved shock pattern. Administration of PSIs (e.g., anisomycin) is modeled by temporarily eliminating the influence of Hebbian plasticity on the weight update.

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): For very short reexposure trials, the shock memory is preferentially retrieved

because it has already been encoded in an attractor as a consequence of acquisition (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 little effect on the shock memory, because after a sufficiently long duration Hebbian learning is operating on a different attractor. Post-reexposure PSI administration has a tangible effect on the shock memory only for intermediate durations (i.e., what we modeled as "short" duration in our simulations of the PSI experiments). In this case, mismatch is large enough to induce degradation of the shock attractor, but not large enough to induce the formation of a new, no-shock attractor. The PSI prevents Hebbian learning from compensating for this degradation by strengthening the shock attractor, so the result is a net decrease in the strength of the shock attractor.

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 intervals 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 acquisition and reexposure. In contrast, our model explicitly represents temporal distance between observations, making it sensitive to changes in timing.[6]

Another, related 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, and no psychological or neurobiological justification was provided. Nonetheless, an 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 we 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 our own, there may be neural implementations of the latent cause model

[6]Conceivably, one could incorporate a time-sensitive mechanism into the Osan model by using a "temporal context" signal that drifts slowly over time (see Sederberg, Gershman, Polyn, & Norman, 2011).

that bring it closer to the formalism of the attractor network. However, in its current form, our model is not specified at the same biologically detailed level as the model of Osan et al. (2011); our model 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 (S. Lee et al., 2008).

Reconsolidation of Episodic Memories

While we have focused on reconsolidation in Pavlovian conditioning, the concept of reconsolidation has recently been used to understand a wider set of experimental findings. In particular, a number of studies have examined putative reconsolidation processes in human episodic memory (J. Chan & LaPaglia, 2013; J. Chan, Thomas, & Bulevich, 2009; Forcato et al., 2007; Forcato, Rodríguez, Pedreira, & Maldonado, 2010; Hupbach, Gomez, Hardt, & Nadel, 2007; Hupbach, Gomez, & Nadel, 2009). In particular, in one line of research developed by Hupbach and colleagues, the researchers used a list-learning paradigm to show that reminding participants of one list (A) shortly before asking them to study a second list (B) 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 (Hupbach et al., 2007). When no reminder was given, participants showed an overall low level of intrusions across list A and list B recall.

One interpretation of these findings, in line with the Monfils-Schiller paradigm, is that the reminder caused the memory of list A to become labile, thereby allowing list B items to become incorporated into the list A memory. Sederberg et al. (2011) showed that the findings of Hupbach and colleagues could be accounted for by a traditional model of memory that does not explicitly incorporate a reconsolidation process (see also Gershman, Schapiro, et al., 2013, for converging neural evidence). It is currently unclear whether similar neurobiological mechanisms underlie reconsolidation of associative and episodic memories.

Conclusion

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

Author note

Samuel J. Gershman, Department of Psychology and Center for Brain Science, Harvard University; Marie-H Monfils, Department of Psychology, University of Texas at Austin; Kenneth A. Norman and Yael Niv, Department of Psychology and Princeton Neuroscience Institute, Princeton University.

This research was supported by a Graduate Research Fellowship from the National Science Foundation (SJG), a Sloan Research Fellowship (YN), and NIMH grants R01MH091147, R21MH086805 (MHM). The authors thank Daniela Schiller for helpful comments.

References

Aggleton, J. P., Sanderson, D. J., & Pearce, J. M. (2007). Structural learning and the hippocampus. *Hippocampus*, *17*, 723–734.

Agren, T., Engman, J., Frick, A., Björkstrand, J., Larsson, E., Furmark, T., & Fredrikson, M. (2012). Disruption of reconsolidation erases a fear memory trace in the human amygdala. *Science*, *337*, 1550–1552.

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

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

Aldous, D. (1985). Exchangeability and related topics. In *École d'Été de probabilités de Saint-Flour xiii* (pp. 1–198). Berlin: Springer.

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

Anderson, J. (1990). *The Adaptive Character of Thought*. Lawrence Erlbaum.

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

Auber, A., Tedesco, V., Jones, C. E., Monfils, M.-H., & Chiamulera, C. (2013). Post-retrieval extinction as reconsolidation interference: methodological issues or boundary conditions? *Psychopharmacology*, *226*, 1–17.

Baker, K. D., McNally, G. P., & Richardson, R. (2013). Memory retrieval before or after extinction reduces recovery of fear in adolescent rats. *Learning & Memory*, *20*, 467–473.

Baldi, E., Lorenzini, C., & Bucherelli, C. (2004). Footshock intensity and generalization in contextual and auditory-cued fear conditioning in the rat. *Neurobiology of Learning and Memory*, *81*, 162–166.

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

Biedenkapp, J., & Rudy, J. (2004). Context memories and reactivation: constraints on the reconsolidation hypothesis. *Behavioral Neuroscience*, *118*, 956–964.

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

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

Blair, H. T., Schafe, G. E., Bauer, E. P., Rodrigues, S. M., & LeDoux, J. E. (2001). Synaptic plasticity in the lateral amygdala: a cellular hypothesis of fear conditioning. *Learning & Memory*, *8*, 229–242.

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

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

Bouton, M. (2004). Context and behavioral processes in extinction. *Learning and Memory*,

11, 485–494.

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

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

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

Capaldi, E. (1994). The sequential view: From rapidly fading stimulus traces to the organization of memory and the abstract concept of number. *Psychonomic Bulletin & Review*, *1*, 156–181.

Chan, J., & LaPaglia, J. A. (2013). Impairing existing declarative memory in humans by disrupting reconsolidation. *Proceedings of the National Academy of Sciences*, *110*, 9309–9313.

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

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

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

Clem, R., & Huganir, R. (2010). Calcium-permeable AMPA receptor dynamics mediate fear memory erasure. *Science*, *330*, 1108–1112.

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

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

Courville, A. (2006). *A latent cause theory of classical conditioning* (Unpublished doctoral dissertation). Carnegie Mellon University, Pittsburgh, PA, USA.

Courville, A., Daw, N., & Touretzky, D. (2006). Bayesian theories of conditioning in a changing world. *Trends in Cognitive Sciences*, *10*(7), 294–300.

Debiec, J., Diaz-Mataix, L., Bush, D., Doyère, V., & LeDoux, J. (2013). The selectivity of aversive memory reconsolidation and extinction processes depends on the initial encoding of the Pavlovian association. *Learning & Memory*, *20*, 695–699.

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

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

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*,

39, 1–38.

Doyère, V., Debiec, J., Monfils, M., Schafe, G., & LeDoux, J. (2007). Synapse-specific reconsolidation of distinct fear memories in the lateral amygdala. *Nature Neuroscience*, *10*, 414–416.

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

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

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

Flavell, C. R., Barber, D. J., & Lee, J. L. (2011). Behavioural memory reconsolidation of food and fear memories. *Nature Communications*, *2*, 504.

Forcato, C., Burgos, V. L., Argibay, P. F., Molina, V. A., Pedreira, M. E., & Maldonado, H. (2007). Reconsolidation of declarative memory in humans. *Learning & Memory*, *14*, 295–303.

Forcato, C., Rodríguez, M. L., Pedreira, M. E., & Maldonado, H. (2010). Reconsolidation in humans opens up declarative memory to the entrance of new information. *Neurobiology of Learning and Memory*, *93*, 77–84.

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

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*, 815–836.

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

Fuhs, M., & Touretzky, D. (2007). Context learning in the rodent hippocampus. *Neural Computation*, *19*, 3173–3215.

Gallistel, C. (2012). Extinction from a rationalist perspective. *Behavioural Processes*, *90*, 66–80.

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

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

Gershman, S., Jones, C. E., Norman, K. A., Monfils, M.-H., & Niv, Y. (2013). Gradual extinction prevents the return of fear: Implications for the discovery of state. *Frontiers in Behavioral Neuroscience*, *7*.

Gershman, S., & Niv, Y. (2010). Learning latent structure: carving nature at its joints. *Current Opinion in Neurobiology*, *20*, 251–256.

Gershman, S., & Niv, Y. (2012). Exploring a latent cause theory of classical conditioning. *Learning & Behavior*, *40*, 255–268.

Gershman, S., & Niv, Y. (2013). Perceptual estimation obeys occam's razor. *Frontiers in Psychology*, *4*.

Gershman, S., Norman, K. A., & Niv, Y. (2015). Discovering latent causes in reinforcement learning. *Current Opinion in Behavioral Sciences*, *5*, 43 - 50.

Gershman, S., Radulescu, A., Norman, K. A., & Niv, Y. (2014). Statistical computations underlying the dynamics of memory updating. *PLOS Computational Biology*, *10*, e1003939.

Gershman, S., Schapiro, A. C., Hupbach, A., & Norman, K. A. (2013). Neural context reinstatement predicts memory misattribution. *The Journal of Neuroscience*, *33*, 8590–8595.

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

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

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

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

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

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

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

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

Ishii, D., Matsuzawa, D., Matsuda, S., Tomizawa, H., Sutoh, C., & Shimizu, E. (2015). An isolated retrieval trial before extinction session does not prevent the return of fear. *Behavioural Brain Research*, *287*, 139–145.

Jarome, T., Kwapis, J., Werner, C., Parsons, R., Gafford, G., & 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*, 300–306.

Jones, C. E., Ringuet, S., & Monfils, M.-H. (2013). Learned together, extinguished apart: reducing fear to complex stimuli. *Learning & Memory*, *20*, 674–685.

Kakade, S., & Dayan, P. (2002). Acquisition and extinction in autoshaping. *Psychological Review*, *109*, 533–544.

Kindt, M., & Soeter, M. (2013). Reconsolidation in a human fear conditioning study: A test of extinction as updating mechanism. *Biological Psychology*, *92*, 43–50.

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

Lattal, K., & 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*, 4667–4672.

Lee, J. (2009). Reconsolidation: maintaining memory relevance. *Trends in Neurosciences*, *32*, 413–420.

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

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

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

Lisman, J., & Grace, A. (2005). The hippocampal-vta loop: controlling the entry of information into long-term memory. *Neuron*, *46*, 703–713.

Ma, X., Zhang, J., & Yu, L. (2012). Post-retrieval extinction training enhances or hinders the extinction of morphine-induced conditioned place preference in rats dependent on the retrieval-extinction interval. *Psychopharmacology*, *221*, 19–26.

McGaugh, J. (2000). Memory–a century of consolidation. *Science*, *287*, 248–251.

McNally, G., Johansen, J., & Blair, H. (2011). Placing prediction into the fear circuit. *Trends in Neurosciences*, *34*, 283–292.

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

Milekic, M., & Alberini, C. (2002). Temporally graded requirement for protein synthesis following memory reactivation. *Neuron*, *36*, 521–525.

Millan, E., Milligan-Saville, J., & McNally, G. P. (2013). Memory retrieval, extinction, and reinstatement of alcohol seeking. *Neurobiology of Learning and Memory*, *101*, 26–32.

Miller, R., & Laborda, M. (2011). Preventing recovery from extinction and relapse. *Current Directions in Psychological Science*, *20*, 325–329.

Miller, R., & Matzel, L. (2006). Retrieval failure versus memory loss in experimental amnesia: Definitions and processes. *Learning & Memory*, *13*, 491–497.

Milton, A., Lee, J., Butler, V., Gardner, R., & 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*, 8230–8237.

Misanin, J., Miller, R., & Lewis, D. (1968). Retrograde amnesia produced by electroconvulsive shock after reactivation of a consolidated memory trace. *Science*, *160*, 554–555.

Monfils, M., Cowansage, K., Klann, E., & LeDoux, J. (2009). Extinction-reconsolidation boundaries: key to persistent attenuation of fear memories. *Science*, *324*, 951–955.

Morris, R., Inglis, J., Ainge, J., Olverman, H., Tulloch, J., Dudai, Y., & Kelly, P. (2006). Memory reconsolidation: sensitivity of spatial memory to inhibition of protein synthesis in dorsal hippocampus during encoding and retrieval. *Neuron*, *50*, 479–489.

Muller, G., & Pilzecker, A. (1900). Experimentelle bietrage zur lehre bom gedachtnisses.

Zeits. Fr Psych., *1*, 1-288.

Nader, K., & Hardt, O. (2009). A single standard for memory: the case for reconsolidation. *Nature Reviews Neuroscience*, *10*, 224–234.

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

Neal, R., & Hinton, G. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models* (pp. 355–368). Cambridge, MA, USA: MIT Press.

Olshavsky, M. E., Jones, C. E., Lee, H. J., & Monfils, M.-H. (2013). Appetitive behavioral traits and stimulus intensity influence maintenance of conditioned fear. *Frontiers in Behavioral Neuroscience*, *7*.

Olshavsky, M. E., Song, B. J., Powell, D. J., Jones, C. E., Monfils, M.-H., & Lee, H. J. (2013). Updating appetitive memory during reconsolidation window: critical role of cue-directed behavior and amygdala central nucleus. *Frontiers in Behavioral Neuroscience*, *7*.

Osan, R., Tort, A., & Amaral, O. (2011). A mismatch-based model for memory reconsolidation and extinction in attractor networks. *PloS One*, *6*, e23113.

Oyarzún, J., Lopez-Barroso, D., Fuentemilla, L., Cucurell, D., Pedraza, C., Rodriguez-Fornells, A., & de Diego-Balaguer, R. (2012). Updating fearful memories with extinction training during reconsolidation: A human study using auditory aversive stimuli. *PloS one*, *7*, e38849.

Pavlov, I. (1927). *Conditioned Reflexes*. Oxford University Press.

Pedreira, M., & Maldonado, H. (2003). Protein synthesis subserves reconsolidation or extinction depending on reminder duration. *Neuron*, *38*, 863–869.

Pedreira, M., Pérez-Cuesta, L., & Maldonado, H. (2004). Mismatch between what is expected and what actually occurs triggers memory reconsolidation or extinction. *Learning & Memory*, *11*(5), 579–585.

Pezze, M., & Feldon, J. (2004). Mesolimbic dopaminergic pathways in fear conditioning. *Progress in neurobiology*, *74*, 301–320.

Power, A., Berlau, D., McGaugh, J., & Steward, O. (2006). Anisomycin infused into the hippocampus fails to block "reconsolidation" but impairs extinction: The role of re-exposure duration. *Learning & Memory*, *13*, 27–34.

Rao-Ruiz, P., Rotaru, D. C., van der Loo, R. J., Mansvelder, H. D., Stiedl, O., Smit, A. B., & Spijker, S. (2011). Retrieval-specific endocytosis of glua2-ampars underlies adaptive reconsolidation of contextual fear. *Nature Neuroscience*, *14*, 1302–1308.

Redish, A., Jensen, S., Johnson, A., & Kurth-Nelson, Z. (2007). Reconciling reinforcement learning models with behavioral extinction and renewal: Implications for addiction, relapse, and problem gambling. *Psychological Review*, *114*, 784–805.

Reichelt, A. C., Exton-McGuinness, M. T., & Lee, J. L. (2013). Ventral tegmental dopamine dysregulation prevents appetitive memory destabilization. *The Journal of Neuroscience*, *33*, 14205–14210.

Rescorla, R. (2004). Spontaneous recovery. *Learning and memory*, *11*, 501–509.

Rescorla, R., & Heth, C. (1975). Reinstatement of fear to an extinguished conditioned stimulus. *Journal of Experimental Psychology: Animal Behavior Processes*, *1*, 88–96.

Rescorla, R., & Wagner, A. R. (1972). A theory of of pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement. In A. Black & W. Prokasy (Eds.), *Classical conditioning ii: Current research and theory* (pp. 64–99). New York, NY: Appleton-Century-Crofts.

Riccio, D., Millin, P., & Bogart, A. (2006). Reconsolidation: A brief history, a retrieval view, and some recent issues. *Learning & Memory*, *13*, 536–544.

Rohrbaugh, M., & Riccio, D. (1970). Paradoxical enhancement of learned fear. *Journal of Abnormal Psychology*, *75*, 210–216.

Rumpel, S., LeDoux, J., Zador, A., & Malinow, R. (2005). Postsynaptic receptor trafficking underlying a form of associative learning. *Science*, *308*, 83–88.

Sanborn, A., Griffiths, T., & Navarro, D. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review*, *117*, 1144–1167.

Sartor, G. C., & Aston-Jones, G. (2013). Post-retrieval extinction attenuates cocaine memories. *Neuropsychopharmacology*, *39*, 1059–1065.

Schiller, D., Kanen, J. W., LeDoux, J. E., Monfils, M.-H., & Phelps, E. A. (2013). Extinction during reconsolidation of threat memory diminishes prefrontal cortex involvement. *Proceedings of the National Academy of Sciences*, *110*, 20040–20045.

Schiller, D., Monfils, M., Raio, C., Johnson, D., LeDoux, J., & Phelps, E. (2010). Preventing the return of fear in humans using reconsolidation update mechanisms. *Nature*, *463*, 49–53.

Sederberg, P., Gershman, S., Polyn, S., & Norman, K. (2011). Human memory reconsolidation can be explained using the temporal context model. *Psychonomic Bulletin & Review*, *18*, 455–468.

Sevenster, D., Beckers, T., & Kindt, M. (2013). Prediction error governs pharmacologically induced amnesia for learned fear. *Science*, *339*, 830–833.

Shors, T., Anderson, M., Curlik, D., & Nokia, M. (2012). Use it or lose it: how neurogenesis keeps the brain fit for learning. *Behavioural Brain Research*, *227*, 450–458.

Soeter, M., & Kindt, M. (2011). Disrupting reconsolidation: pharmacological and behavioral manipulations. *Learning & Memory*, *18*, 357–366.

Soto, F. A., Gershman, S. J., & Niv, Y. (2014). Explaining compound generalization in associative and causal learning through rational principles of dimensional generalization. *Psychological Review*, *121*, 526-558.

Spear, N. (1973). Retrieval of memory in animals. *Psychological Review*, *80*, 163–194.

Squire, L. (2006). Lost forever or temporarily misplaced? the long debate about the nature of memory impairment. *Learning & Memory*, *13*, 522–529.

Steinfurth, E. C., Kanen, J. W., Raio, C. M., Clem, R. L., Huganir, R. L., & Phelps, E. A. (2014). Young and old Pavlovian fear memories can be modified with extinction training during reconsolidation in humans. *Learning & Memory*, *21*, 338–341.

Sutton, R., & Barto, A. (1998). *Reinforcement learning: An introduction*. Cambridge University Press.

Suzuki, A., Josselyn, S., Frankland, P., Masushige, S., Silva, A., & Kida, S. (2004). Memory reconsolidation and extinction have distinct temporal and biochemical signatures. *Journal of Neuroscience*, *24*, 4787–4795.

Tronson, N., & Taylor, J. (2007). Molecular mechanisms of memory reconsolidation. *Nature Reviews Neuroscience*, *8*, 262–275.

Wang, L., & Dunson, D. (2011). Fast Bayesian inference in Dirichlet process mixture models. *Journal of Computational and Graphical Statistics*, *20*, 196–216.

Wang, S., de Oliveira Alvares, L., & Nader, K. (2009). Cellular and systems mechanisms of memory strength as a constraint on auditory fear reconsolidation. *Nature Neuroscience*, *12*, 905–912.

Widrow, B., & Hoff, M. (1960). Adaptive switching circuits. *1960 WESCON Convention Record Part IV*, 96–104.

Winters, B., Tucci, M., & DaCosta-Furtado, M. (2009). Older and stronger object memories are selectively destabilized by reactivation in the presence of new information. *Learning & Memory*, *16*, 545–553.

Xue, Y.-X., Luo, Y.-X., Wu, P., Shi, H.-S., Xue, L.-F., Chen, C., ... others (2012). A memory retrieval-extinction procedure to prevent drug craving and relapse. *Science*, *336*, 241–245.

Zhu, X., Ghahramani, Z., & Lafferty, J. (2005). Time-sensitive Dirichlet process mixture models. *Technical Report, CMU-CALD-05-104, Carnegie Mellon University*.

Appendix: Computational Model Details

In this section, we provide the mathematical and implementational details of our model.

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 our 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

$$\begin{aligned}\mathcal{F}(\mathbf{W}, Q) &= \sum_{\mathbf{z}_{1:t}} Q(\mathbf{z}_{1:t}|\mathcal{D}_{1:t}) \log P(\mathbf{z}_{1:t}, \mathcal{D}_{1:t}|\mathbf{W}) \\ &= \sum_{\mathbf{z}_{1:t}} Q(\mathbf{z}_{1:t}|\mathcal{D}_{1:t}) \log \left[P(\mathcal{D}_{1:t}|\mathbf{z}_{1:t}, \mathbf{W}) P(\mathbf{z}_{1:t})\right]. \end{aligned} \tag{11}$$

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 & 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,

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

Alternating the E and M steps repeatedly, $\mathcal{F}(\mathbf{W}, Q)$ is guaranteed to converge to a local maximum (Neal & 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 we summarize an approximate, online form of the algorithm. To reduce notational clutter, we drop the n superscript (indicating EM iteration), and implicitly condition on $\mathbf{W}$.

The E-step: Structure Learning

The E-step corresponds to calculating the posterior using Bayes' rule:

$$q_{tk} = P(z_t = k|\mathcal{D}_{1:t}) = \frac{\sum_{\mathbf{z}_{1:t-1}} P(\mathcal{D}_t|z_t = k, \mathcal{D}_{1:t-1}) P(z_t = k|\mathbf{z}_{1:t-1})}{\sum_j \sum_{\mathbf{z}_{1:t-1}} P(\mathcal{D}_t|z_t = j, \mathcal{D}_{1:t-1}) P(z_t = j|\mathbf{z}_{1:t-1})}. \tag{12}$$

Note that 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), we use a "local" *maximum a posteriori* (MAP) approximation (see Sanborn et al., 2010, for more discussion):

$$q_{tk} \approx \frac{P(\mathcal{D}_t|z_t = k, \hat{\mathbf{z}}_{1:t-1}, \mathcal{D}_{1:t-1}) P(z_t = k|\hat{\mathbf{z}}_{1:t-1})}{\sum_j P(\mathcal{D}_t|z_t = j, \hat{\mathbf{z}}_{1:t-1}, \mathcal{D}_{1:t-1}) P(z_t = j|\hat{\mathbf{z}}_{1:t-1})}, \tag{13}$$

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

$$\hat{z}_t = \operatorname*{argmax}_k P(\mathcal{D}_t|z_t = k, \hat{\mathbf{z}}_{1:t-1}, \mathcal{D}_{1:t-1}) P(z_t = k|\hat{\mathbf{z}}_{1:t-1}). \tag{14}$$

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 our simulations it tends to produce very similar solutions to more elaborate approximations like particle filtering (Gershman & Niv, 2010; Sanborn et al., 2010).[7]

The first term in Eq. 14 (the likelihood) is derived using standard results in Bayesian statistics (Bishop, 2006):

$$P(\mathcal{D}_t|z_t = k, \hat{\mathbf{z}}_{1:t-1}, \mathcal{D}_{1:t-1}) = \mathcal{N}(r_t; \hat{r}_{tk}, \sigma_r^2) \prod_{d=1}^{D} \mathcal{N}(x_{td}; \hat{x}_{tkd}, \nu_{tk}^2), \tag{15}$$

where

$$\hat{r}_{tk} = \sum_{d=1}^{D} x_{td} w_{kd} \tag{16}$$

$$\hat{x}_{tkd} = \frac{N_{tk} \bar{x}_{tkd}}{N_{tk} + \sigma_x^2} \tag{17}$$

$$\nu_{tk}^2 = \frac{\sigma_x^2}{N_{tk} + \sigma_x^2} + \sigma_x^2. \tag{18}$$

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. 14 (the prior) is given by the time-sensitive Chinese restaurant process (Eq. 3).

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

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:

$$[\nabla \mathcal{F}]_{kd} = \sigma_r^{-2} x_{td} \delta_{tk}, \tag{19}$$

where δ_{tk} is given by Eq. 6. To make the similarity to the Rescorla-Wagner model clearer, we absorb the σ_r^{-2} factor into the learning rate, η.

Simulation Parameters

With two exceptions, we 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, we treated θ and λ as free parameters, which we fit using least-squares. For simulations of the human data in Figure 11, we 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, our results do not depend strongly on these particular values.

Relationship to the Rescorla-Wagner Model

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

$$\hat{r}_t = \sum_{d=1}^{D} w_d x_{td} \tag{20}$$

$$\delta_t = r_t - \hat{r}_t \tag{21}$$

$$\mathbf{w} \leftarrow \mathbf{w} + \eta \mathbf{x}_t \delta_t. \tag{22}$$

The key difference is that in our model, we 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.