ORIGINAL ARTICLE

Journal Section

Explaining the effectiveness of fear extinction through latent-cause inference

Mingyu Song[1] | Carolyn E. Jones[2,3,4] | Marie-H. Monfils[4] | Yael Niv[1,5]

[1] Princeton Neuroscience Institute, Princeton University, Princeton, NJ 08540

[2] Department of Neurology, Oregon Health & Science University, Portland, OR 97239

[3] Research & Development, VA Portland Healthcare System, Portland, OR 97239

[4] Department of Psychology and Institute for Mental Health Research, The University of Texas at Austin, Austin, TX 78712

[5] Department of Psychology, Princeton University, Princeton, NJ 08540

Correspondence
Mingyu Song
Email: mingyus@princeton.edu

Funding information
MS and YN were supported by the National Institute of Drug Abuse (R01DA042065).

Acquiring fear responses to predictors of aversive outcomes is crucial for survival. At the same time, it is important to be able to modify such associations when they are maladaptive, for instance in treating anxiety and trauma-related disorders. Standard extinction procedures can reduce fear temporarily, but with sufficient delay or with reminders of the aversive experience, fear often returns. The latent-cause inference framework explains the return of fear by presuming that animals learn a rich model of the environment, in which the standard extinction procedure triggers the inference of a new latent cause, preventing the unlearning of the original aversive associations. This computational framework had previously inspired an alternative extinction paradigm – gradual extinction – which indeed was shown to be more effective in reducing the return of fear. However, the original framework was not sufficient to explain the pattern of results seen in the experiments. Here, we propose a formal model to explain the effectiveness of gradual extinction in reducing spontaneous recovery and reinstatement effects, in contrast to the ineffectiveness of standard extinction and a gradual reverse control procedure. We demonstrate through quantitative simulation that our model can explain qualitative behavioral differences across different extinction procedures as seen in the empirical study. We verify the necessity of several key assumptions added to the latent-cause framework, which suggest potential general principles of animal learning and provide novel predictions for future experiments.

KEYWORDS
fear extinction, latent cause inference, gradual extinction

1 | INTRODUCTION

Fear memories are notoriously hard to erase. After an association has been formed between an originally neutral cue (e.g., a tone) and some aversive outcome (e.g., a foot shock), animals cannot simply unlearn this association through standard extinction procedures (i.e., being presented with the cue repeatedly in the absence of the aversive outcome) [1]. During extinction, the animals' fear response towards the cue gradually reduces, but it usually returns if tested after a long delay (spontaneous recovery)[2, 3], or if the animal is reminded of the aversive outcome (reinstatement)[2, 4]. Associative learning theory explains these phenomena by postulating that extinction involves learning a *new* association rather than updating the original fear association [5]. However, this theory does not delineate the particular circumstances under which a new association is initiated, and how this can be avoided so that the original association may be modified through new experience.

Recent theoretical work [6] suggested a formalization of the way animals decide when to learn a new association and when to update old associations using a *latent-cause inference* framework. In this framework, all observations (e.g., cues and reinforcers) are assumed to be generated by latent (unobservable) causes that are each active for a certain (unknown) amount of time. Each latent cause has a certain tendency to generate observations, characterized by its "generative strength"[1] of each observation. While interacting with the environment, animals infer what latent cause is currently active based on their prior experience and current observations, and behave accordingly. At the same time, they learn and update their estimates of the generative strength of the currently active latent cause.

According to this latent-cause inference framework, in fear extinction, animals infer that there are two distinct latent causes, based on their distinct tendency to generate shocks: one dangerous latent cause (active during conditioning, with a high probability of generating shocks), and one safe latent cause (active during extinction, with low or no probability of generating shocks). The reduction of fear response during extinction is a result of the animal's increasing belief that the second cause is active as more no-shock observations accumulate. Presentation of a shock in a reinstatement procedure is taken to indicate that the original dangerous cause is likely to be active again, causing the return of fear. Similarly, after some passage of time, both causes (dangerous and safe) are equally likely to be active again, leading to the spontaneous recovery of fear response as compared to that measured shortly after extinction.

In addition to explaining the return of fear, the latent-cause inference framework prescribes a solution for effective extinction of the original fear association: instead of abruptly "cutting off" the pairing between tone and shock,

[1] In associative learning, the tendency for two stimuli to co-occur is characterized by an "associative strength". Inspired by this, here we term the tendency for a latent cause to generate observations "generative strength" to emphasize the causal/generative relationship. This concept is also similar to the "emission probability" in Hidden Markov Models.

which encourages the animals to infer a new latent cause, gradually decreasing their co-occurrence will make animals more likely to infer that the old cause is still active during extinction, but with decreasing tendency to generate shocks. To demonstrate this, Gershman and colleagues [7] conducted two *gradual extinction* experiments (see Figure 1 for experimental design), and indeed they observed reduced fear responses in both a spontaneous recovery test and a reinstatement test. In both experiments, they contrasted this condition with both a standard extinction condition and a *gradual reverse* condition. In the latter, instead of gradually decreasing the frequency of the shock, they gradually increased it, while keeping the total number of shocks the same. Despite the surface similarity between the two gradual conditions (only two trials were different, though this difference markedly changed the trend of shock frequency from decreasing to increasing; Figure 2), extinction was far less successful in preventing the return of fear in the gradual reverse condition, supporting the idea that abrupt changes encourage the inference of new latent causes. Since then, the effectiveness of gradual extinction has also been replicated in human participants [8]. There is also evidence of effective extinction in other experimental paradigms that create a gradual transition between the consistent presentation of shocks (during conditioning) and the absence of shocks. For example, occasional reinforcements during extinction eliminated spontaneous recovery in humans [9, 10], and replacing the strong footshock used in conditioning with weak shocks during extinction (i.e., deconditioning) helped reduce both renewal and spontaneous recovery effects in rodents [11].

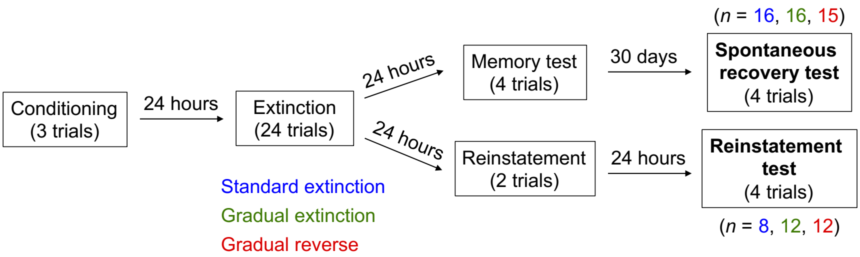


FIGURE 1 Experimental design in Gershman et al. [7]. Across two experiments (spontaneous recovery and reinstatement), rats were assigned to three different extinction conditions: standard extinction, gradual extinction and gradual reverse. All animals first underwent a conditioning session (3 trials of tone-shock pairing). This was followed by an extinction session (procedures differed based on the extinction condition) 24 hours later. In the spontaneous recovery experiment, animals were first tested on their memory of extinction after another 24 hours (termed a "long-term memory test" in the original paper), and then tested for spontaneous recovery of fear response 30 days later. In the reinstatement experiment, animals underwent 2 reinstatement trials (shocks presented alone) 24 hours after extinction, and then were tested for fear response after another 24 hours. The number of animals participating in each experimental condition is noted for each experiment, color-coded based on the extinction procedure: standard extinction in blue, gradual extinction in green, and gradual reverse in red.

Conceptually, these effects can be explained by the latent-cause inference framework [6]; in fact, the experiments done by Gershman and colleagues [7] were inspired by this theoretical framework and aimed to test its predictions. However, it turned out that the original latent-cause inference model (presented in detail, e.g., in [12]), was not able to explain the empirical observations, in particular the difference between gradual extinction and gradual reverse conditions (S. Gershman, personal communication). The lack of a satisfactory computational account limited the conclusions that could be drawn from the original work, and its potential to inspire future investigation.

To address this gap, here we describe a latent-cause model with additional assumptions that captures the pattern of the original empirical results. We show the model's predictions through simulations, and demonstrate the necessity

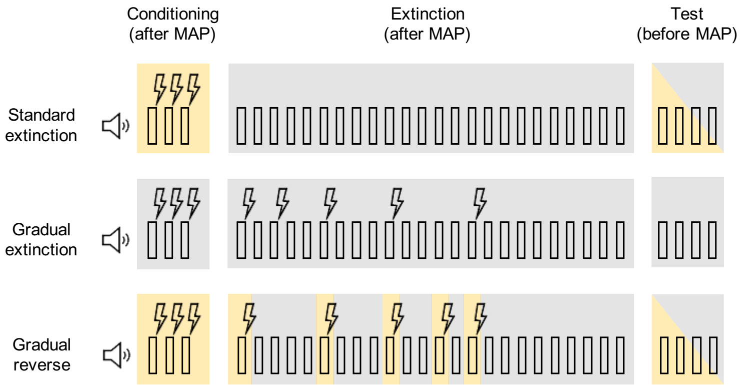


FIGURE 2 Trial sequences and model predictions for latent-cause assignments under the three extinction conditions. Trial sequence in conditioning, extinction and test sessions: each rectangle represents the tone presentation in one trial; trials in which the tone co-terminated with a shock are marked by the lightning signs. Background shading colors indicate latent-cause assignments predicted by our model: a two-cause sequence for standard extinction and gradual reverse, and a one-cause sequence for gradual extinction. Splitting color at test indicates that both causes are likely. The possibility of a third, new, latent cause at test is not illustrated, as it is largely consistent across conditions and contributes little to freezing behavior. For conditioning and extinction sessions, the latent-cause assignments shown are those after each session, after collapsing to the mode of the posterior (MAP estimation; see text). For the test session, we show the probabilistic assignments before the MAP estimation as those generated the behavior measured. That is, all predicted assignments shown (both for conditioning and extinction sessions, and for the test session) are presumably what animals had in mind at test time, and therefore what governed freezing behavior during test.

of each model assumption by comparing with alternative models. Indeed, it is not trivial for the model to generate predictions that match animals' behavior across all three extinction conditions, and several important modeling assumptions are needed. These include assumptions on animals' prior belief over latent causes, the way in which they learn about the statistics of each latent cause, how animals reduce uncertainty in their inference process, and animals' response tendencies. We note that even with these assumptions, our model can only predict the comparative behavioral patterns between conditions, but not the exact freezing rates in each condition. We discuss potential reasons for this discrepancy. Although the original empirical results we model here may benefit from additional examination to show its generality, we propose that understanding the potential computational underpinnings of these findings can advance our understanding of extinction processes and how they can be made more effective. More broadly, this work provides insights into the potential mechanisms that support animal learning and inference, and generates new predictions to be tested in future experiments.

2 | METHODS

2.1 | The latent cause inference model

We use a latent-cause inference model to explain the effectiveness (in terms of degree of return-of-fear) of standard extinction, gradual extinction and gradual reverse procedures, in both spontaneous recovery and reinstatement experiments. The model describes how animals infer the active latent cause on each trial by combining a prior belief with current observations, and how animals learn about the statistics of each latent cause. We assume such inference and learning were active throughout the entire experiment (including test sessions; see Results for experiment details). Specifically, as detailed below, the model we found to account for the experimental results uses the distance-dependent Chinese restaurant process as the prior, the Rescorla-Wagner update rule to learn about "generative strength" (i.e., the likelihood for each latent cause to generate observations, as detailed below), and assumes that animals approximate the (intractable) Bayesian inference process by collapsing their belief to the posterior mode between sessions. Additionally, to compare model predictions to empirical measurements, we make assumptions about how the prediction of shock maps to freezing behavior. We now describe each part of the model.

2.1.1 | Prior: distance-dependent Chinese restaurant process

We assume that the animals' prior belief over latent causes takes the form of a distance-dependent Chinese restaurant process [13], a variant of the Chinese restaurant process (CRP) infinite mixture-model prior [14].

The standard CRP describes a categorization process with an *a priori* unlimited number of categories, whereby a new trial is more likely to be generated by a latent cause (category) that has generated more trials in the past. Specifically, the prior probability of an old cause generating the current trial is proportional to the number of trials this cause has already generated; the probability of the next trial being generated by a completely new latent cause is proportional to a fixed concentration parameter α. Denoting the active latent cause on trial i by c_i, the prior probability distribution over c_i is thus:

$$P(c_i = c | c_{1:i-1}) = \begin{cases} \frac{1}{i-1+\alpha} \sum_{j<i} \delta(c_j, c) & (c \text{ is an old cause}) \\ \frac{1}{i-1+\alpha}\ \alpha & (c \text{ is a new cause}) \end{cases}$$

where $\delta(x, y)$ is the Kronecker delta function: $\delta(x, y) = 1$ if $x = y$; otherwise, $\delta(x, y) = 0$. Thus, $\delta(c_j, c)$ denotes whether the current cause c is the same one that generated trial j. $\frac{1}{i-1+\alpha}$ is the normalization constant for this distribution. This distribution is exchangeable, meaning that it results in the same prior distribution over latent causes regardless of the order of trials.

Because trial order is important in the task we model, we used the the distance-dependent Chinese restaurant process (ddCRP), in which more distant experience contributes less to current inference through a decay function applied to the trial count. Specifically, we compute distance over time, using an exponential function with slope k. Since exponential decay can be arbitrarily close to zero (that is, an old latent cause not experienced for a long time can have a close-to-zero prior probability), departing from the classic ddCRP, we also add a baseline probability b to all old latent causes. This corresponds to an assumption that any old latent cause has a non-negligible prior probability of becoming active again. The equations then become:

$$P(c_i = c | c_{1:i-1}) \propto \begin{cases} \sum_{j<i} e^{-k(t_i - t_j)} \delta(c_j, c) + b & (c \text{ is an old cause}) \\ \alpha & (c \text{ is a new cause}) \end{cases} \tag{1}$$

where t_i is the time of trial i. The normalization constant for this distribution can be calculated by summing over the probability of all old causes and the new cause.

2.1.2 | Likelihood: Rescorla-Wagner learning

We denote the generative strength of each observation x for latent cause c by $V(x|c)$, which describes the tendency of latent cause c to generate x. Using the simplest and most widely-used learning rule in animal conditioning, i.e., the Rescorla-Wagner learning rule [15], we assume that V is updated for the currently active latent cause c_i based on the observation of x:

$$\Delta V(x|c_i) = \eta(x_i - V(x|c_i))$$

where η is the learning rate, and x_i is the observation of x on trial i ($x_i = 1$ or 0 means x is present or absent, respectively, on that trial).

To model fear extinction, we consider two types of observations: tone and shock. Given the different *a priori* prevalence of such stimuli in the animals' natural environment, we assume different initial values for the generative strengths for new latent causes: $V_0(\text{tone}) = 0.5$ and $V_0(\text{shock}) = 0.05$. We also assume a higher learning rate η_{shock} for shocks (when shocks are present) considering their high motivational valence. These assumptions are important for the pattern of results, but the specific numeric values were not chosen through formal optimization or model-fitting.

We use the generative strength V as the proxy for the animal's estimated probability of observing the tone or shock, i.e., the likelihood of the corresponding observation on that trial given a latent cause:

$$P(x|c_i) = V(x|c_i). \tag{2}$$

2.1.3 | Update: Exact inference, with collapse of belief distribution between sessions

We assume that animals perform Bayesian inference during each session. That is, they combine the prior probability and likelihood of both observations (assumed to be independently generated given the latent cause) to calculate the posterior belief distribution over the active latent cause:

$$P(c_i|\mathbf{x}, c_{1:i-1}) \propto P(c_i|c_{1:i-1}) \prod_{x \in \mathbf{x}} P(x|c_i).$$

Here, the first term on the right-hand side is the ddCRP prior from above (Equation 1) and the second term is the likelihood of the current observations (Equation 2 above). **x** denotes the combination of both observations (tone and shock). In this way, the animal maintains a probability distribution over each of the latent causes being active on each trial, updating this distribution as trials unfold.

Between experimental sessions, however, we assume that animals collapse their posterior belief distribution to its mode, i.e., a maximum *a posteriori* (MAP) estimation. In other words, we assume that animals do not maintain uncertainty over what latent cause was responsible for what observation in the previous session; instead, they "pick" the most likely sequence of latent causes for the past session, moving forward to the next session with only this deterministic assignment of trials to latent causes as the prior. Note that this is not a technical choice for faster model simulation, but an important modeling assumption for predicting the difference between gradual extinction and gradual reverse conditions (see 3.4.3).

2.1.4 | Mapping prediction of shock to freezing behavior

On each trial, as animals hear the tone and before the observation of a shock (on reinforced trials), we assume that animals use their current estimate of latent-cause assignment to predict how likely a shock is to occur on that trial, and thus decide whether or not to freeze in anticipation. The estimated shock probability is calculated by marginalizing over all possible latent causes:

$$P(\text{shock}|\text{tone}) = \sum_{c_i} P(\text{shock}|c_i)P(c_i|\text{tone}, c_{1:i-1}).$$

Here, the last term is calculated using Bayes rule:

$$P(c_i|\text{tone}, c_{1:i-1}) \propto P(c_i|c_{1:i-1})P(\text{tone}|c_i).$$

In mapping the animal's prediction of shock probability to freezing behavior, we make two more assumptions. First, we assume a non-zero baseline freezing rate (denoted by p_0): animals do not freeze in their natural environments; if, however, the animal becomes aware of shocks through experimental manipulations and anticipates them, empirical findings suggest that animals will show some baseline level of freezing behavior [16], regardless of how unlikely the shock is. In addition, for simplicity, we assume that freezing probability is proportional to the predicted shock probability:

$$P(\text{freezing}) = (1 - p_0) * P(\text{shock}|\text{tone}) + p_0.$$

We also consider the locally perseverative nature of animals' behavior [17, 18], and assume that there is a chance of p_r that the animal will exhibit the same behavior as in last trial, regardless of its current prediction for the presence or absence of shock.

2.2 | Model simulations

We used the following parameter values for simulating the model: $\alpha = 0.2, k = 0.1, b = 0.1, \eta = 0.2, \eta_{\text{shock}} = 0.4, V_0(\text{tone}) = 0.5, V_0(\text{shock}) = 0.05, p_0 = 0.2, p_r = 0.7$. These values were initially set based on parameter values used in similar previous studies (α, η) or a best first guess given our knowledge of animal behavior and the role of each parameter in the model ($k, b, \eta_{\text{shock}}, V_0(\text{tone}), V_0(\text{shock}), p_0, p_r$). We then fine-tuned the parameters with a grid search around that starting point, to best match empirical results. We did not perform likelihood-optimizing model-fitting because it is computationally prohibitive for this model. However, we note that simulation results were consistent across a range of parameter values. We also explored how model predictions depend on the change in parameter values (Supplementary Figure 1). Because it is computationally intractable to compute the full posterior distribution analytically, we used the particle filter algorithm as in [6] to approximate the posterior distribution with 10,000 particles.

We simulated the model and compared its predictions with the empirical results for the main experiments [7] (in Results) and a few additional experiments (in Discussion). The experimental protocols used in these experiments and our simulations, are described in detail in the following sections.

3 | RESULTS

In the following, we describe behavioral predictions of the model with all the assumptions described above (distance-dependent prior on cause assignment, Rescorla-Wagner rule for learning the generative strength of stimuli, MAP estimation between sessions, direct mapping from shock prediction to freezing behavior), and compare them to behavioral results in Gershman et al. [7]. We then turn to evaluating the necessity of each assumption by comparing with alternative models.

3.1 | Experimental measures and modeling goals

We simulated the model and obtained its prediction for the three extinction conditions: standard extinction, gradual extinction, and gradual reverse. We considered two types of test (as in [7]): spontaneous recovery and reinstatement. Here, we first describe the experimental conditions in brief (see [7] for details on experimental design and subjects), as well as the goals of our modeling.

In the experiment, each rat completed one of the experimental conditions (Figure 1). All experiments started with a conditioning session (3 trials of a 20s tone, co-terminated with a 0.5s foot-shock). 24 hours later, animals underwent an extinction session (24 trials of the same tone) in one of three ways (Figure 2): in standard extinction, the tone was presented alone in all trials; in gradual extinction, the tone co-terminated with a shock on trials 1, 3, 6, 10 and 15, and no shocks on other trials; in gradual reverse, the tone co-terminated with a shock on trials 1, 6, 10, 13 and 15, and no shocks on other trials. In this way, in gradual extinction, shocks gradually became less frequent, whereas in gradual reverse, shocks were made to be more frequent as the extinction session proceeded. In all cases, the last 9 trials of the extinction session did not involve shocks, to allow comparable extinction of freezing behavior before the subsequent test.

Animals were then tested for their fear response to the tone, measured by the percentage of time they spent freezing during the tone. In the spontaneous recovery experiment, 24 hours after the extinction session, the animals first underwent a so-called long-term memory test to test the extinction memory (with 4 trials of the tone alone); 30 days later, they were tested again for spontaneous recovery (4 trials of the tone alone). In the reinstatement experiment, 24 hours after the extinction session, animals experienced 2 unsignaled shocks without the tone. After another 24 hours, they were tested with 4 trials of the tone alone.

Our simulations focused on model predictions for (1) latent-cause assignment throughout the experiment; (2) animals' freezing rate in the last 4 test trials. Our goals were to demonstrate the latent-cause assignments that support animals' behavior under each extinction procedure, and more importantly, explain the qualitative differences between the three procedures in terms of their effectiveness in preventing the return-of-fear: gradual extinction being the most effective (i.e., lowest freezing rate at test, compared to the end of extinction), in comparison to standard extinction and gradual reverse. We note here, and return to this point in the Discussion, that while we strove to predict behavior throughout the experiment, due to large variability between individual rats, the behavioral pattern during extinction was hard to discern. We therefore did not attempt to quantitatively predict trial-by-trial behavior, or to fit the behavior of individual animals by using different parameter values for each animal. Instead, we focused on the average qualitative pattern of results at test, which meaningfully separated the different extinction procedures in terms of effectiveness.

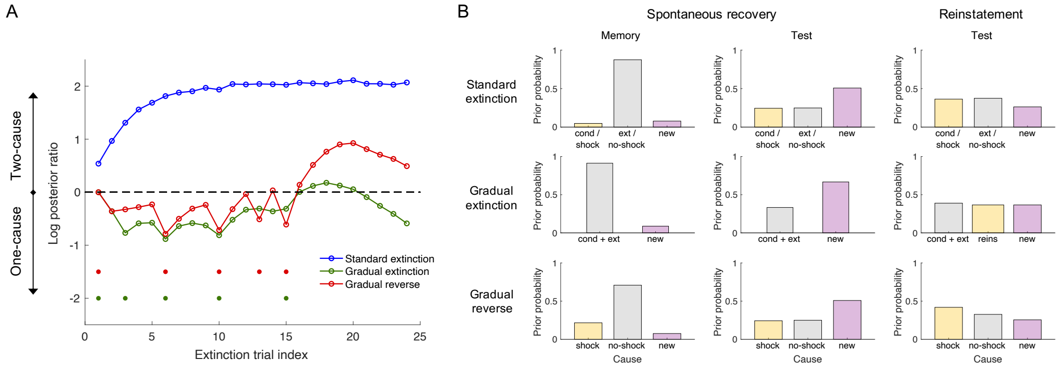


FIGURE 3 Model prediction for latent-cause probabilities. (A) Comparison between a two-cause sequence and one-cause sequence during extinction. Log posterior ratio greater than (less than) zero indicates the dominance of the two-cause (one-cause) sequence in inference; the dashed line at zero represents the two types of cause sequence being equally likely. At the end of the extinction session, the two-cause sequence is more likely than the one-cause sequence in standard extinction and gradual reverse; in contrast, the one-cause sequence is more likely in gradual extinction. Here, "two-cause" sequence corresponds to the assignment of all shock trials to one cause, and all non-shock trials to another; "one-cause" sequence corresponds to all conditioning and extinction trials being generated by the same cause. **(B) Prior probability of latent causes in the first trial of the long-term memory and test sessions**. Left and middle columns: long-term memory and test sessions in the spontaneous recovery experiment; right column: test session in the reinstatement experiment. Top, middle and bottom rows: standard extinction, gradual extinction, and gradual reverse conditions, respectively. Causes are labeled based on what types of past trials they have generated: conditioning (cond), extinction (ext), reinstatement (reins), shock or no-shock. They are color-coded as in Figure 2: yellow indicates a "dangerous cause", grey indicates a "safe cause", and purple indicates a new cause (with minimal prediction of shock, $V_0(\text{shock}) = 0.05$).

3.2 | Latent-cause assignment

Figure 2 illustrates the model's assignment of trials to latent causes for the three extinction conditions. In this schematic, assignments in conditioning and extinction sessions are the deterministic results after the post-session MAP estimation (see Figure 3A for the probabilistic assignments during the extinction session); these are the assignments that influence behavior at the test session, which is the focus of our interest. We also combine here the two types of test (spontaneous recovery and reinstatement) because their latent-cause assignments are similar (see Figure 3B for latent-cause probability in each test). In standard extinction, two different latent causes are inferred for the conditioning session (all shock trials) and the extinction session (all no-shock trials); at test, both causes are likely (due to either reminder shocks or passage of time, both elevating the probability of the initial conditioning-session latent cause). In gradual extinction, because of the gradual reduction of shock probability, conditioning and extinction sessions are assigned to the same latent cause, as are the test trials. In gradual reverse, in contrast, all shock trials throughout conditioning and extinction are assigned to one latent cause, whereas all no-shock trials are assigned to a different latent cause; then, at test, both causes are likely.

3.3 | Prediction of freezing behavior

Figure 4 shows model prediction for freezing rate across experiment sessions. For all three extinction conditions, freezing rate increases during conditioning, decreases during extinction (more so in standard extinction, with no shocks

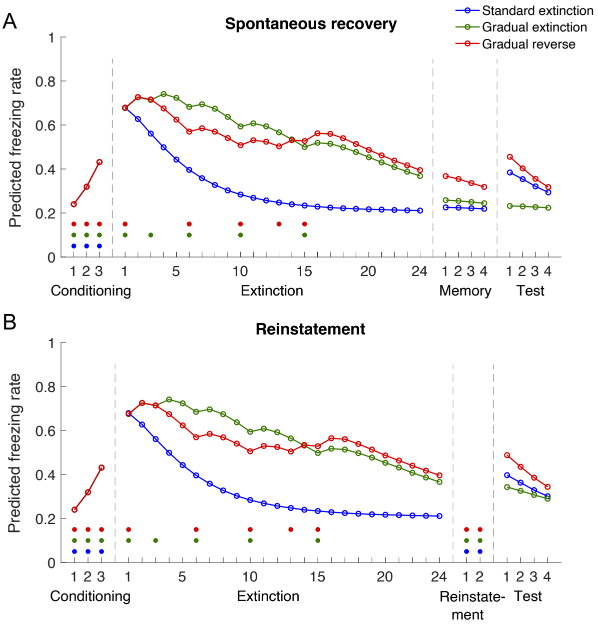


FIGURE 4 Model prediction of freezing rate for (A) spontaneous recovery and (B) reinstatement experiments. Under all three extinction conditions, freezing rate increases during conditioning, and decreases during extinction. In the beginning of test session, however, freezing rate jumps back up for standard extinction and gradual reverse conditions, but remains low in gradual extinction. Note that the model predicts freezing rate upon tone presentation, before the actual delivery (or absence) of shock. Dots at the bottom of the plots indicate shocks in the corresponding trials, color-coded based on the extinction conditions. Dashed gray vertical lines indicate session boundaries (in practice: at least 24h gap).

in the extinction session), and continues to decrease in the long-term memory test (24 hours after extinction in the spontaneous recovery experiment). However, in the test session (both spontaneous recovery and reinstatement experiments), the predictions for the three conditions diverge, due to the distinct latent-cause assignments: in standard extinction and gradual reverse conditions, freezing rate increases compared to the end of extinction, showing the return of fear; in the gradual extinction condition, it continues to decrease both in the spontaneous recovery and the reinstatement tests.

Figure 5A and 5B summarize model predictions on the difference in freezing rate between the four test trials and the last four extinction trials, for comparison with the qualitative pattern in the empirical results (Figure 5C and 5D). According to the model, for both spontaneous recovery and reinstatement tests, fear response reduces the most in gradual extinction, followed by gradual reverse; fear response at test increases in standard extinction. We note here (and discuss in more detail below) the clear discrepancies between our simulation results and the empirical results, where in simulation gradual reverse does not result in overall increase in fear at test as compared to the end of extinction. Nevertheless, the model predictions are qualitatively consistent with the empirical findings, illustrating the success of the current model in explaining the relative pattern seen in the experimental results.

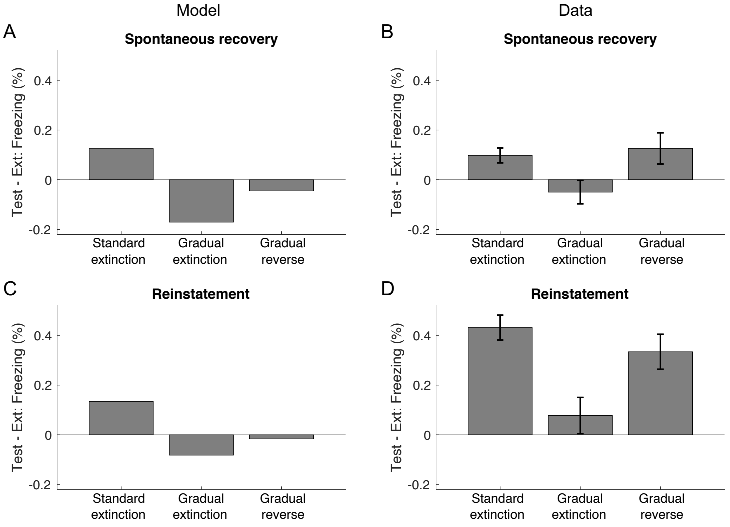


FIGURE 5 Impact of different forms of extinction on return of fear: model predictions (A,C) are qualitatively consistent with empirical results (B,D) in both experiments. Model simulations correctly predict that standard extinction leads to the greatest return of fear, whereas gradual extinction is the most effective in permanently reducing fear, across both spontaneous recovery and reinstatement tests. The effectiveness of extinction is calculated as the difference in freezing rate between the four test trials and the last four extinction trials (all no-shock trials). Panels B and D are reproduced from [7].

3.4 | Necessity of model assumptions

The model described above was tailored to explain the behavioral patterns in the data by adding assumptions as needed where the original CRP model [6] did not suffice. We now turn to discussing these model assumptions and demonstrating their necessity. We do so by comparing the predictions of the current model (referred to as “main” model below) to reduced models that do not include these assumptions. Since we did not fit the parameters of the main model to the data using statistical techniques, we do not present formal model comparisons, but rather focus on the qualitative patterns in the data.

3.4.1 | Distance-dependent CRP: spontaneous recovery depends on test delay

Experiments suggest that spontaneous recovery of fear response increases with the delay between extinction and test [2, 3]. For example, the standard extinction group in Gershman et al. [7] showed no significant return of fear in the memory test 24 hours after extinction (paired sample t-test: $t(15) = -1.80, p = .09$, in comparison to the last four extinction trials). In contrast, the rats showed significantly more freezing at test 30 days later, compared to the last four trials of extinction (paired sample t-test: $t(15) = 3.26, p < .01$).

Our model captures the dependency of spontaneous recovery on delay duration using the distance-dependent CPR prior, similar to what was used in [19]. According to this prior, the probability of an old latent cause being active depends on the time between its previous instances and the current trial. The closer they are in time, the more likely the old cause will be active again. As a result, the model predicts that the latent cause inferred for the extinction

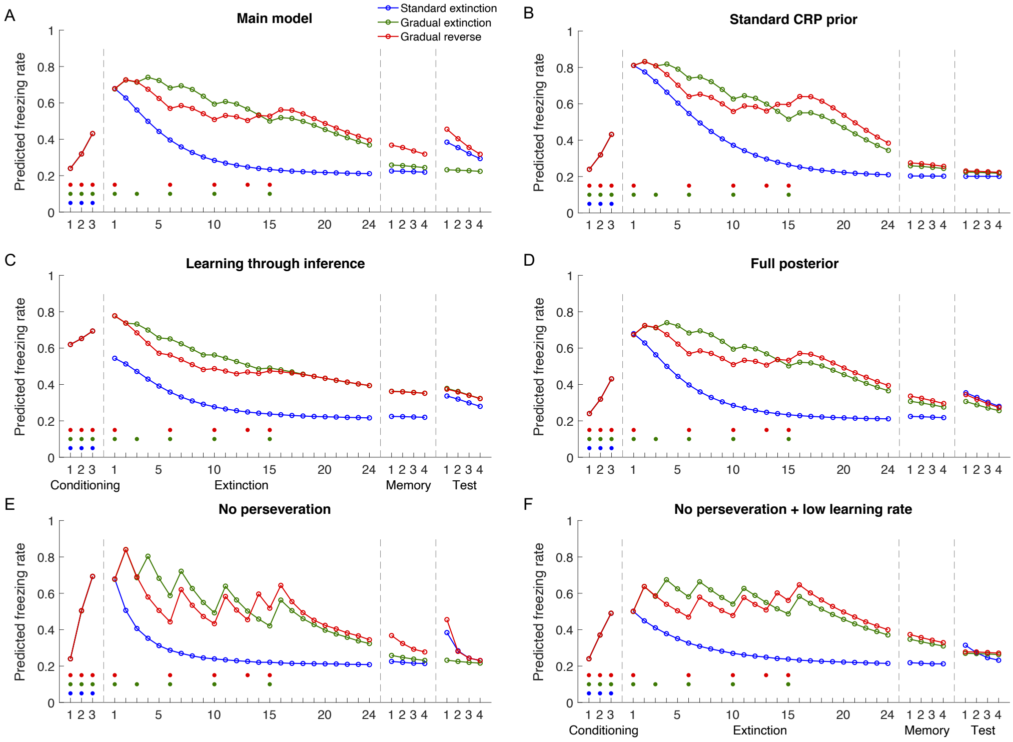


FIGURE 6 Simulation of alternative models demonstrates the necessity of assumptions in the main model. Each alternative model differs from the main model in only one assumption; model simulations were conducted with the same parameter values, except for the alternative assumptions, as noted below. Shown are simulations of the spontaneous recovery experiment; results are consistent for the reinstatement experiment. **(A)** The main model, same as Figure 4A. **(B)** Standard CRP prior (without distance dependence; $k = 0$ in the ddCRP) assigns trials in the memory test (24 hours later) and spontaneous recovery test (30 days later) to the same latent causes, and thus shows no spontaneous return-of-fear after 30 days for any of the conditions. **(C)** Learning generative strengths through inference (i.e., Bayesian inference on Bernoulli probabilities for tone and shock, instead of Rescorla-Wagner learning) predicts the same latent-cause assignments and expected shock probability for gradual extinction and gradual reverse conditions, and thus cannot explain their difference in fear responses during test. In this alternative model, we replaced the Rescorla-Wagner learning parameters $\eta, \eta_{\text{shock}}, V_0(\text{tone}), V_0(\text{shock})$ with parameters for pseudo-counts: $N(\text{tone}) = 0.5, N(\text{no-tone}) = 0.5, N(\text{shock}) = 0.95, N(\text{no-shock}) = 0.05$. These pseudo-counts determine the prior probabilities of tone and shock (each denoted by x): $p(x) = N(\text{x})/(N(\text{x}) + N(\text{no-x}))$, reflecting a similar asymmetry in the *a priori* prevalence of tone and shock as in the main model. **(D)** Keeping the full posterior distribution over latent causes across sessions (no MAP estimation) predicts minimal return-of-fear in the gradual reverse condition. **(E)** No perseveration ($p_r = 0$) leads to over-sensitivity to shock/no-shock experience during extinction which was not seen in the empirical data, as well as rapid reduction in return-of-fear during test. **(F)** Replacing the perseveration assumption with lower learning rates to stabilize responding across trials ($p_r = 0, \eta = 0.1, \eta_{\text{shock}} = 0.2$) fails to predict the return-of-fear effect in the gradual reverse condition.

session is very likely to continue to be active in the memory test (Figure 3B left column), as the memory test is much closer to the extinction session (24 hours apart) than to the conditioning session (48 hours). After 30 days, however, both conditioning and extinction sessions are similarly distant, so the model predicts that both the conditioning and extinction latent causes are equally likely to be active (Figure 3B middle column; of course, a completely new latent cause is more likely in this case), resulting in an increased fear response in standard extinction and gradual reverse. In gradual extinction, since there has been only one latent cause throughout, this cause (with reduced generative strength of shock) is inferred to be active again during test, resulting in no return-of-fear above and beyond what was observed at the end of extinction.

An alternative model that uses the standard CRP prior with no distance dependence (and keeps all other assumptions and parameter values the same as the main model) produced latent-cause assignments that are independent of test time. It thus made similar predictions for the memory test and the spontaneous recovery test, failing to predict the spontaneous recovery of fear in standard extinction and gradual reverse conditions (Figure 6B).

Note that we are not committed here to the specific form of the distance dependence and have not exhaustively tested other forms of recency-weighted or perseverative CRP priors (e.g., [20]), or different decay functions. Our claim is only that an order-agnostic exchangeable prior distribution, as is commonly used in machine learning applications to categorization, would not accord with the empirical data.

3.4.2 | Rescorla-Wagner learning (recency-weighted estimates): difference between gradual extinction and gradual reverse

In the latent-cause inference framework, it is common to assume that the generative strength of observations is fixed (though unknown) for each latent cause [6, 12]. Under this assumption, the optimal estimator of the shock probability for each latent cause is the proportion of trials in which shocks appeared under that cause. This assumption is at odds with the behavioral differences between gradual extinction and gradual reverse, as both conditions had the same number of shocks and would therefore predict identical test behavior (Figure 6C).

To explain the behavioral difference between gradual extinction and gradual reverse, we thus found it necessary to assume a recency-weighted estimate of generative strength, such as given by the Rescorla-Wagner (RW) learning rule. Similar learning rules have been used in latent-cause inference models of associative learning and memory modification [21, 19]. The RW rule learns a dynamic shock probability through an error-correcting process that adjusts estimates proportionally to the error in predicting the current observation. A fixed learning rate results in over-weighing of recent experiences, effectively estimating the generative strength as an exponentially-decreasing average of previous experiences. This is normative if animals assume that the environment may change over time and track such changes. In our case, this allows them to treat differently a decrease in shock probability in gradual extinction versus an increase in gradual reverse. As a result, they make different latent cause assignments for gradual extinction and gradual reverse, and in turn, show distinct levels of fear responses at test.

3.4.3 | Collapsing uncertainty (MAP) between sessions: return-of-fear in gradual reverse

Latent-cause inference involves evaluating the likelihood of all possible cause sequences for the past trials – from all trials being generated by one cause, to each trial being generated by its own unique cause, through any combination in between. Probabilistic inference means that the model does not commit to any of these assignment sequences; instead, all are likely during inference, and each is associated with some non-zero probability. In particular, in all three extinction conditions, both a one-cause sequence and a two-cause sequence are likely (Figure 3A). While the two-cause sequence dominates in standard extinction, the assignments for gradual extinction and gradual reverse are more uncertain, starting with similar probabilities in both conditions and diverging through extinction. At the end of the extinction session, the one-cause sequence establishes some advantage over the two-cause sequence for gradual extinction, and the opposite is true for gradual reverse. The MAP assumption collapses this uncertainty over latent-cause sequences at the end of the extinction session, obtaining the deterministic (and different) assignments shown in Figure 2. Specifically, the model commits to the one-cause assignment for gradual extinction, and the two-cause assignment for gradual reverse, accentuating the small differences between these conditions at the end of the extinction session. Similar collapsing of uncertainty has been applied in domains like perceptual decision-making (e.g. [22]) to facilitate inference and decisions where estimating the full posterior distribution is computationally intractable.

We found the MAP assumption to be necessary for predicting the return-of-fear effect in gradual reverse. An alternative model that keeps the full distribution over latent causes throughout the experiment failed to predict the increase in freezing rate during test for gradual reverse (Figure 6D). This is because the two-cause sequence was deemed only somewhat more likely during test, which resulted in minimal return of fear. Similarly, because of the probabilistic latent cause assignments, the two-cause sequence had a non-zero probability in gradual extinction, resulting in a slight increase of freezing rate at test compared to the main model prediction.

3.4.4 | Perseveration: gradual change in behavior during extinction, and persistence of return-of-fear during test

In the main model, we postulate that animals tend to repeat what they have been doing in past trials with probability p_r. Such perseverative behavior has been widely observed in perceptual and value-based choice tasks, for both animals and humans [23, 24, 25, 26]. In the current task, perseveration is important for predicting the persistence of return-of-fear during the test session. Without this assumption, fear responses will decrease rapidly in both standard extinction and gradual reverse conditions, as early as the second test trial (Figure 6E). This is due to animals' inference of latent causes: experiencing one trial without shock is sufficient to infer that the "safe cause" is active in the test session, and due to the distance-dependent CRP prior, this inference comes to dominate latent-cause assignments at test. However, such rapid reduction in freezing behavior was not seen behaviorally.

Simulations without perseveration also predicted freezing behavior that was overly sensitive to shock and no-shock experiences during extinction (Figure 6E): in gradual extinction and gradual reverse, simulated freezing rate jumped up following shocks and dropped following no-shocks. Examining animals' freezing rate during extinction (by re-scoring of the original videos using a convolutional neural network [27], as the middle trials during extinction session were not scored or reported in the original paper) suggested that behavior did not reflect trial-by-trial shock delivery or absence; on average, freezing rate noisily but gradually decreased through extinction session in all three conditions (not shown). An alternative explanation for the more gradual behavioral change in extinction can be a lower learning rate in updating the generative strengths of latent causes. However, reducing the learning rate interfered with latent-cause inference as it rendered animals less sensitive to shocks and thus biased them towards the one-cause assignment. An alternative model where we halved the learning rates predicted no return-of-fear in gradual reverse (Figure 6F).

We therefore used the perseveration assumption to account for both rapid learning in fear conditioning and extinction, and gradual change in behavior. Additionally, we note that due to local perseveration, the main model predicted an increase in freezing rate at the beginning of the extinction session (compared to the end of conditioning session; Figure 6A), potentially related to the well-documented "extinction burst" phenomenon [28, 29].

4 | DISCUSSION

In this work, we use a latent-cause inference model to explain the differential effectiveness of gradual versus abrupt (standard) fear-extinction procedures. The model explains the return-of-fear effect commonly observed in standard extinction by presuming that animals infer a new state of the world during extinction and thus form a new association between tone and shock, as opposed to unlearning the original association. Similar ideas have been proposed both under similar statistical inference framework [30, 6] and through reinforcement-learning models with a state-classification mechanism [31]. This explanation also aligns with decades-old suggestions that extinction results in learning of a new safe association that competes with the original threat association, though does not override it [5]. The novelty of our work lies in formulating the inference and learning processes, demonstrating with quantitative simulation the differences between three extinction procedures, and verifying the necessity of various model assumptions.

We show through model simulations that animals make distinct inferences of latent cause assignments under different extinction procedures: gradual extinction is the most effective at extinguishing the original fear association because both conditioning and extinction sessions are assigned to the same latent cause, and the gradual reduction in shock helps animals acquire a decreasing estimate of shock probability, leading to minimal return-of-fear during test. In contrast, the abrupt change in shock frequency in the other two procedures (complete absence of shocks in standard extinction; abrupt reduction in shock appearance, followed by increasing shock frequency in gradual reverse) results in the creation of a new “safe” latent cause to which all subsequent no-shock trials are assigned. This new latent cause protects the old “dangerous” latent cause from being updated by the no-shock experiences. As a result, the original fear association remains intact and can resurface at test.

4.1 | Additional model assumptions

We found that several additional model assumptions were needed to predict the behavioral findings: (1) a distance-dependent prior on latent-cause assignments that used the passage of time to determine distance; (2) learning the dynamics of the environment through a recency-weighted rule such as the Rescorla-Wagner learning rule; (3) reduction of uncertainty over the posterior distribution through MAP estimation between sessions; (4) behavioral perseveration. Without each of these assumptions, the model failed to replicate the higher freezing rates at test in standard extinction and gradual reverse conditions in comparison to gradual extinction. Most of these assumptions build on past models that have successfully explained a wide range of phenomena in animal (and human) learning and decision-making, as we discuss below. Moreover, the necessity of each assumption also suggests principles of animal learning mechanisms in Pavlovian tasks and beyond.

The distance-dependent CRP prior we used highlights the important role of time in latent-cause inference, especially when behavior is examined at different time intervals, or even on different time scales. We are not the first to propose such time-dependency. Gershman and colleagues [19] used a similar distance-dependent CRP prior (with a power-law temporal kernel) to explain how spontaneous recovery depends on extinction-test interval, as well as why the effect of post-retrieval memory modification is sensitive to memory age. Additional empirical evidence for time-dependent inference process comes from work on extinction delay. Myers and colleagues [32] found that varying the delay between acquisition and extinction affected the amount of return-of-fear: animals that experienced extinction trials 10 minutes or 1 hour after acquisition showed little or no subsequent spontaneous recovery, reinstatement or renewal effects; whereas those who had extinction trials 24 or 72 hours after acquisition showed strong return of fear effects. This effect of extinction delay cannot be explained by a latent-cause inference model without temporal dependency. The distance-dependent CRP prior can explain these findings as a result of the decreasing probability of the extinction trials being generated by the same latent cause as the acquisition trials when the delay between extinction and acquisition increases. Thus, with a shorter delay (10 min or 1h), the animal is more likely to classify the extinction trials into the same latent cause as the acquisition trials, which helps the successful unlearning of the shock probability, and thus prevents the return of fear.

The importance of time is also reflected in the recency-weighted learning rule we used. The Rescorla-Wagner learning rule has been widely used to model classical conditioning. It has also been used in latent-cause inference models to explain compound generalization in associative and causal learning [21] and memory modification [19].

Together, the distance-dependent CRP prior and the Rescorla-Wagner learning rule imply specific beliefs that animals may have about the environment. Both modeling assumptions reflect the animal's inner model of the environment, i.e., whether they consider it as static or changing over time. According to the distance-dependent CRP prior, latent causes that were last active a long time ago are less likely to be active again, suggesting that the passage of time can lead to changes in the environment, making older causes less likely. Similarly, the Rescorla-Wagner learning rule over-weighs more recent experience, inherently capturing how the statistics of observations within a latent cause may change over time. Alternative assumptions (standard CRP prior and learning through exact Bayesian inference) are better suited to static environments. The advantage of the current model over alternatives provides evidence that animals are capable of acquiring rich dynamics in their learning environment, and that their inner model of the environment is in accord with the actual changing nature of naturalistic environments.

We can further consider a more general approach to modeling learning in a changing environment: deriving the normative learning rule based on the generative model. For example, the Kalman filter model [33] has been suggested as a model for estimating the mean and standard deviation of a Gaussian reward distribution under the assumption that it evolves over time[2]. Another option, normative for environments in which the amount of reward changes at a constant rate, is to add a "momentum" term (calculated as the running average of recent prediction errors) to the Rescorla-Wagner learning rule [35].[3] Future work can derive normative learning rules for the three extinction procedures in the current experiment, and test whether they account for behavior better than the Rescorla-Wagner learning rule we used.

Compared to the above two assumptions, the MAP assumption (collapse of the posterior to its mode between sessions) is found less often in the animal-learning literature. This assumption, nevertheless, also reflects the effect of the passage of time on inference, and can be construed as a result of memory consolidation (e.g., via replay of past events during sleep [36]). Through consolidation, animals may revisit their learning experience in the previous session, and continue to update their belief over latent causes accordingly. Such update can reinforce the most probable latent-cause assignment, and eventually make it the only possibility, equivalent to a MAP estimate. Similar consolidation mechanisms have been proposed during long inter-trial intervals. For example, to explain memory modification, Gershman and colleagues [19] introduced a "rumination" process taking place between the re-exposure of previous memory and the attempt to modify/extinguish it. Such rumination reinforces the dominant belief of returning to the past context, and facilitates memory modification. There is also extensive evidence on the superiority of spaced learning (with longer intervals between training examples, including overnight session boundaries) over massed learning, facilitated by memory consolidation [37], in both animals and humans [38, 39, 40, 41]. Specifically, Urcelay and colleagues [42] found that spaced extinction trials (3 trials per session, for 3 sessions in total) led to less spontaneous recovery and renewal of the original memory than massed trials (9 trials in a single session). We replicated this result

[2] In fact, Rescorla-Wagner learning can be seen as a special case of the Kalman filter with a fixed Kalman gain and without tracking uncertainty [34]

[3] We also tested an alternative model with momentum; the results were largely consistent with the RW learning rule. Thus, for simplicity, we used the more basic RW rule in the main model.

by simulating our model (albeit with a lower value of the concentration parameter α than used in other simulations; Figure 7A). In the model, the difference between the conditions is caused by the collapse of the posterior distribution between sessions: over a single long session, the inferred probability for the two-cause sequence increases throughout the session, eventually exceeding that of the one-cause sequence and leading to a two-cause belief. In contrast, at the end of each short session of spaced extinction, the one-cause sequence is still more probable and is therefore "committed" through the MAP mechanism. This results in a one-cause belief at the end of extinction, and therefore animals show less return of fear under spaced extinction. Similar results have also been found in deconditioning experiments [11] where a single long deconditioning session leads to return-of-fear but multiple short ones do not.

From a normative perspective, MAP estimation may be justified to allow inference in complex scenarios where the full distribution over latent causes is computationally intractable. Because of the combinatorial explosion of possible latent-cause sequences as more trials are experienced, any experience that unfolds over a sequence of events quickly becomes intractable. It is thus reasonable to assume that the brain collapses uncertainty over previous inference periodically (perhaps facilitated by large gaps in experience, replay of experiences, and memory consolidation), to be able to continue building on past knowledge without necessarily maintaining all of it. Indeed, Stocker and colleagues [22, 43] have shown that biases in human decision making can be explained by postulating that people collapse parts of their posterior distribution between sequential decisions, essentially discarding beliefs that are inconsistent with actions they have already made. Our MAP assumption is a similar form of commitment to the most likely past beliefs over others.

In a different but related neural network model of memory modification [44, 11], attractors are used to model memory traces, similar to the latent causes in the current framework. The probabilistic retrieval and modification of existing attractors, and the formation of new attractors are similar to the inference and learning on latent causes. In the network model, learning results in stored attractors in the network, which is conceptually similarly to the deterministic latent-cause assignments as a result of MAP approximation in the current model.

The MAP assumption provides novel theoretical predictions that can be tested in future experiments. A direct test of the memory consolidation account of the MAP estimation would be to reduce the time between extinction and test in the reinstatement experiment. For example, conducting extinction, reinstatement and test all within one day or even one session. If memory consolidation indeed facilitates the collapse of the posterior distribution and strengthening of one interpretation of past events, animals with no or less time between extinction and test should maintain the probabilistic belief at test, and thus show less freezing in the gradual reverse condition. Our model does not differentiate between an abrupt collapse of the posterior distribution and a gradual reduction of posterior uncertainty. To further examine the process of uncertainty reduction, future work can manipulate the time interval between extinction and test, and examine how freezing rate at test changes as a function of such time interval.

The reduction of uncertainty due to the MAP estimation may seem at odds with the increase in uncertainty over time that results from the dynamic generative model as discussed earlier. However, it is worth noting that the MAP reduction of uncertainty affects categorization of past experiences, whereas the increase in uncertainty pertain to predicting future experience. On the one hand, the collapse of the posterior may suggest the limitation of animals' representation of the world; perhaps animals are able to learn a rich representation, but fail to maintain uncertainty about this representation over long periods of time. On the other hand, MAP estimation can also be seen as a method by which animals (and humans) build concise models of the world.

Last but not least, the perseveration assumption suggests a value-free habitual system (i.e., animals tend to repeat their previous behavior regardless of their prediction in the current scenario, as posited by Miller and colleagues [17]) that exists alongside a model-based system corresponding to the latent cause inference mechanism in the current model. Such a dichotomy has been widely observed in animal learning and decision-making. The fact that animals'

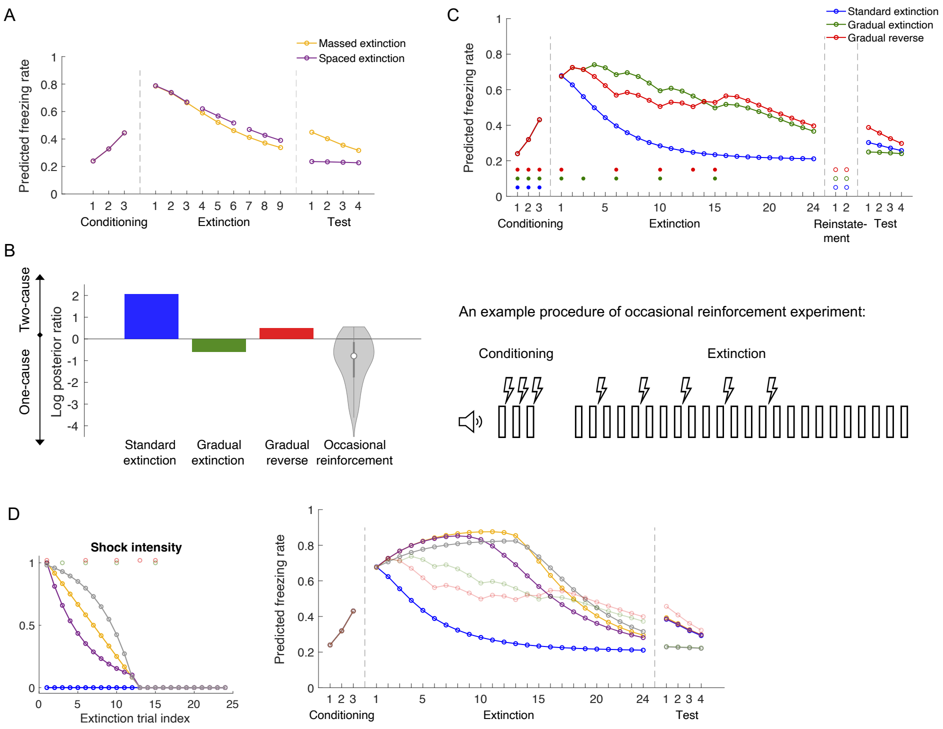


FIGURE 7 Simulations of additional experiments. The same model and parameter values were used as in the main experiments, unless otherwise noted. **(A)** Spaced extinction is better than massed extinction at reducing long-term fear. We simulated the spaced and massed extinction procedures similar to [42], with spaced extinction comprised of 3 trials per session for 3 sessions in total and massed extinction comprised of 9 trials in a single session (all extinction trials involved presentation of the tone without a shock). Spontaneous recovery was tested after 30 days, and showed considerable return of fear for the massed condition, and markedly less fear in the spaced condition. In this simulation, $\alpha = 0.05$. Note that the lower α value used here indicated a lower tendency to infer new latent causes, which could be due to the animals' additional prior shaping experience in [42]. **(B)** Occasional reinforcement during extinction favors the inference of one latent cause over two latent causes. To simulate the occasional reinforcement, five shocks were randomly placed in the first 15 trials in the extinction session, with the terminal 9 trials being no-shock trials; 100 different experiments (each with shocks randomly allocated to trials) were simulated. Left: log posterior ratio between a two-cause belief and a one-cause belief at the end of extinction session, for the three extinction procedures as in [7] and the occasional reinforcement procedure. Violin plot shows the median (dot), interquartile range (darker error bar), and 95% confidence interval (lighter error bar) across the 100 simulations. Right: an example procedure of the occasional reinforcement experiment. **(C)** Low intensity reinstatement (modeled as no-shock trials) leads to minimal return of fear. The same experimental setup was used as in [7], except that there was no shock in the two reinstatement trials (to approximate animals' perception of low intensity shocks). There was less return of fear at test compared to the original results (Figure 4B) in both standard extinction and gradual reverse procedures. **(D)** Gradually reducing shock intensity prevents return-of-fear. Left: shock intensity in several extinction procedures with reduced intensity over time (in gray, yellow, and purple curves; note that the three curves overlap at 0 in trials 13 to 24.). The three procedures reported in [7] are also shown as references (standard extinction: blue; gradual extinction: green; gradual reverse: red; all shocks in gradual extinction and gradual reverse were modeled with an intensity of 1, plotted with jitters for better visualization; no-shock trials were modeled with an intensity of 0, not plotted). Right: predicted freezing rate during conditioning, extinction and spontaneous recovery test (30 days later); color codes as in the left panel (note that blue, purple and yellow curves overlap at test). The procedure with the most gradual change in shock intensity at the beginning of extinction (gray curve, in contrast to yellow and purple curves) best prevents long-term return of fear. In this simulation, shock intensity was modeled as a continuous variable with Gaussian noise $N(0, \sigma^2)$ (σ is an additional model parameter which was set to 0.43; all other parameter values were kept as in the original model).

behavior does not fully reflect their learned world model or even stimulus values underscores the importance of accounting for common habitual behavior (e.g., perseveration, side-bias, etc.) in modeling so that the habitual part of behavior will not mask the rich learning mechanisms.

4.2 | Related empirical results

In this work, we provided a quantitative and theoretical account explaining why gradual extinction is more effective in permanently reducing fear than standard extinction and gradual reverse procedures. We focused specifically on predicting the empirical findings in Gershman et al. [7]; however, it is worth noting that there is other evidence on the effectiveness of gradual extinction.

In a fear-extinction experiment with human participants [8], gradual extinction was shown to prevent the return of fear better than standard extinction, as measured by startle response (although there were no effects for contingency rating or skin-conductance response). Additional evidence comes from occasional reinforcement experiments: having occasional reinforced trials during extinction (effectively reducing the shock probability more gradually compared to standard extinction procedure) has been shown to eliminate spontaneous recovery in humans [9, 10]. Similarly, in appetitive conditioning experiments, occasional presentations of reinforcement in extinction slowed the re-acquisition of conditioned responses [45, 46], suggesting unlearning during extinction rather than learning of a competing association that would allow rapid relearning by activating the original association. We simulated the effect of receiving occasional reinforcement during extinction with our model (Figure 7B): according to the model, by the end of extinction, animals would be more likely to believe that conditioning and extinction sessions were generated from the same cause, rather than two distinct causes; therefore, they would show less return of fear.

We note that seemingly opposing evidence was observed in a reinstatement experiment by Rescorla [4], where gradual extinction was less effective than standard extinction in preventing the reinstatement effect. However, the reminder shocks used in this experiment were very weak (compared to other reinstatement experiments reported in the same study: 0.5mA vs 3mA), and therefore may not have been perceived as aversive stimuli by the animals [47]. Without valid reminder shocks, the test trials would simply reflect the effect of extinction after a short delay (similar to the long-term memory test in the spontaneous recovery experiment in Gershman et al. [7]). Indeed, the gradual extinction group showed a higher conditioned response than the standard extinction group in the memory test in Gershman et al. [7] ($t(30) = 3.05, p < .005$), consistent with the findings by Rescorla [4]. In fact, simulating our model with low-intensity reinstatement trials showed minimal return of fear in subsequent tests across all three procedures (Figure 7C). Given the binary definition of shocks in our model, we modeled the low-intensity trials as no-shock trials, which could reflect how animals perceived such shocks due to the nonlinear psychophysics of pain perception [48].

Our model simulation suggests that an alternative way to reduce long-term fear memory is to gradually reduce the intensity of shocks (Figure 7D). Similar to the gradual extinction procedure used in [7], the gradual reduction of intensity can facilitate the inference of one single latent cause. Specifically, the model predicts that the change in shock intensity needs to be gradual in the beginning of the extinction session (gray curve in Figure 7D); otherwise, animals will still show return of fear (yellow and purple curves). According to our model prediction, this new procedure is similarly effective as the original gradual-extinction procedure (green curve). However, a quantitative comparison depends on the values of model parameters. Empirical work is needed to test the new procedure and compare it with the original gradual extinction procedure.

4.3 | Limitations: differences between model predictions and empirical results

Although the current model captures key aspects of the empirical findings, its predictions deviate from the experimental data quantitatively. First, the model correctly predicts the comparative differences between the three extinction conditions, but fails to predict the absolute return-of-fear effects. It over-predicts the reduction of fear for gradual extinction and gradual reverse, as compared to the empirical results (Figure 5). In fact, the model predicts lower fear responses in gradual extinction at test, and no change in gradual reverse when compared to the end of extinction, whereas animals showed minimal change and an increase in fear in these two conditions, respectively. Similarly, the predicted freezing rate during extinction deviates from the empirical results quantitatively (Supplementary Figure 2).

However, it is worth noting that behavioral variability was profound in these experiments. For instance, despite having the exact same procedures in conditioning and extinction sessions, the extinction effects differed in spontaneous recovery and reinstatement experiments (Supplementary Figure 2). Specifically, during the last four trials of extinction, freezing rate was significantly different between the two experiments under the same standard extinction procedure (one-way ANOVA: $F(1,22) = 4.38, p < .05$; there were no significant difference for gradual extinction or gradual reverse procedures). These potentially different freezing rates at the end of extinction, nevertheless, served as the baselines for comparing test behavior in Figure 5, and provided the “ground truth” for comparison with our simulation results. Given this variability across animals and between experiments in the empirical findings, we decided to forego matching the empirical results quantitatively, and instead focused on correctly predicting the comparative difference between extinction procedures: gradual extinction being the most effective in reducing the return of fear, compared to either standard extinction or gradual reverse.

Quantitative deviations between the model’s predictions and the empirical results may also be due to non-linear mapping between estimated shock probability and animals’ freezing behavior, which we did not model. Because the form of this mapping was not the focus of the current work, we assumed a linear function for simplicity, but this is likely incorrect. Future work can design targeted experiments to investigate the underlying mechanism of this mapping.

Finally, the model captures behavior at the group level but does not make predictions regarding individual differences, which were marked in the empirical results (also observed in similar studies in humans [49]). To make individual predictions with the current model, we can use different parameter values for each animal to capture their distinct learning, inference and behavioral processes. For example, animals often demonstrate abrupt switching (rather than gradual changes) in behavior that only seem gradual at the group level because different animals switch behavior at different times [50]. This between-animal variability in change points may be the result of different mapping functions from shock prediction to freezing rate, and may be an alternative to the perseveration assumption in explaining the gradual change of behavior over time. Another possibility is that individual animals do not perform full Bayesian inference but only take a small number of noisy samples from the probability distribution; in this case, aggregating over a group of animals will result in average behavior that resembles full inference, all the while each individual shows what seems like an idiosyncratic pattern of behavior [51]. Under this explanation, the current model is only adequate for explaining group-level behavior, and we need an additional sampling model to capture individual behavior. Both accounts for individual differences are likely in the task we modeled. We leave for future work to examine these sources of variability by fitting models to individual animals’ behavior and comparing their predictions.

4.4 | Conclusion

In sum, our work explains the effectiveness, or lack thereof, of different behavioral manipulations aimed at reducing maladaptive fear responses. Our results suggest that in acquiring and extinguishing fear responses, animals form a dynamic model of the environment, using inference of latent causes to predict future events. When representing and memorizing past experiences, however, our model suggests that animals summarize previous inference by collapsing distributions over potential latent-cause explanations and preserving the most likely explanation. Our findings suggest that, even in simple Pavlovian tasks such as fear extinction, animals' behavior can reveal a rich array of mechanisms of learning and representation.

Acknowledgements

The authors would like to thank Lili X. Cai for assistance with the scoring neural network and helpful discussions, and Samuel J. Gershman for helpful comments on the draft. The contents of this paper do not represent the views of the U.S. Department of Veterans Affairs or the United States Government.

Data and Code Availability

All data and code used in this paper are available at `https://github.com/mingyus/fear-extinction-latent-cause-inference`.

References

[1] Dunsmoor JE, Niv Y, Daw N, Phelps EA. Rethinking extinction. Neuron 2015;88(1):47–63.

[2] Pavlov IP. Conditioned reflexes: an investigation of the physiological activity of the cerebral cortex. Oxford, England: Oxford Univ. Press; 1927.

[3] Rescorla RA. Spontaneous recovery. Learning & Memory 2004;11(5):501–509.

[4] Rescorla RA, Heth CD. Reinstatement of fear to an extinguished conditioned stimulus. Journal of Experimental Psychology: Animal Behavior Processes 1975;1(1):88.

[5] Bouton ME. Context and behavioral processes in extinction. Learning & memory 2004;11(5):485–494.

[6] Gershman SJ, Blei DM, Niv Y. Context, learning, and extinction. Psychological review 2010;117(1):197.

[7] Gershman SJ, Jones CE, Norman KA, Monfils MH, Niv Y. Gradual extinction prevents the return of fear: implications for the discovery of state. Frontiers in behavioral neuroscience 2013;7:164.

[8] Shiban Y, Wittmann J, Weißinger M, Mühlberger A. Gradual extinction reduces reinstatement. Frontiers in behavioral neuroscience 2015;9:254.

[9] Thompson A, McEvoy PM, Lipp OV. Enhancing extinction learning: Occasional presentations of the unconditioned stimulus during extinction eliminate spontaneous recovery, but not necessarily reacquisition of fear. Behaviour Research and Therapy 2018;108:29–39.

[10] Culver NC, Stevens S, Fanselow MS, Craske MG. Building physiological toughness: Some aversive events during extinction may attenuate return of fear. Journal of behavior therapy and experimental psychiatry 2018;58:18–28.

[11] Popik B, Amorim FE, Amaral OB, Alvares LDO. Shifting from fear to safety through deconditioning-update. Elife 2020;9:e51207.

[12] Gershman SJ, Niv Y. Exploring a latent cause theory of classical conditioning. Learning & behavior 2012;40(3):255–268.

[13] Blei DM, Frazier PI. Distance Dependent Chinese Restaurant Processes. Journal of Machine Learning Research 2011;12(8).

[14] Aldous DJ. Exchangeability and related topics. In: École d'Été de Probabilités de Saint-Flour XIII—1983 Springer; 1985.p. 1–198.

[15] Rescorla RA. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. Current research and theory 1972;p. 64–99.

[16] Jacobs NS, Cushman JD, Fanselow MS. The accurate measurement of fear memory in Pavlovian conditioning: resolving the baseline issue. Journal of neuroscience methods 2010;190(2):235–239.

[17] Miller KJ, Shenhav A, Ludvig EA. Habits without values. Psychological review 2019;126(2):292.

[18] Lau B, Glimcher PW. Dynamic response-by-response models of matching behavior in rhesus monkeys. Journal of the experimental analysis of behavior 2005;84(3):555–579.

[19] Gershman SJ, Monfils MH, Norman KA, Niv Y. The computational nature of memory modification. Elife 2017;6:e23763.

[20] Lloyd K, Leslie DS. Context-dependent decision-making: a simple Bayesian model. Journal of The Royal Society Interface 2013;10(82):20130069.

[21] Soto FA, Gershman SJ, Niv Y. Explaining compound generalization in associative and causal learning through rational principles of dimensional generalization. Psychological review 2014;121(3):526.

[22] Stocker AA, Simoncelli EP. A Bayesian model of conditioned perception. Advances in neural information processing systems 2007;2007:1409.

[23] Ito M, Doya K. Validation of decision-making models and analysis of decision variables in the rat basal ganglia. Journal of Neuroscience 2009;29(31):9861–9874.

[24] Gold JI, Law CT, Connolly P, Bennur S. The relative influences of priors and sensory evidence on an oculomotor decision variable during perceptual learning. Journal of neurophysiology 2008;100(5):2653–2668.

[25] Lee D, McGreevy BP, Barraclough DJ. Learning and decision making in monkeys during a rock–paper–scissors game. Cognitive Brain Research 2005;25(2):416–430.

[26] Akaishi R, Umeda K, Nagase A, Sakai K. Autonomous mechanism of internal choice estimate underlies decision inertia. Neuron 2014;81(1):195–206.

[27] Cai LX, Pizano K, Gundersen GW, Hayes CL, Fleming WT, Holt S, et al. Distinct signals in medial and lateral VTA dopamine neurons modulate fear extinction at different times. Elife 2020;9:e54936.

[28] Skinner BF. The behavior of organisms: an experimental analysis. Appleton-Century; 1938.

[29] Cooper JO, Heron TE, Heward WL. Applied behavior analysis. Pearson Education, Incorporated; 2020.

[30] Courville AC, Daw ND, Touretzky DS. Bayesian theories of conditioning in a changing world. Trends in cognitive sciences 2006;10(7):294–300.

[31] Redish AD, Jensen S, Johnson A, Kurth-Nelson Z. Reconciling reinforcement learning models with behavioral extinction and renewal: implications for addiction, relapse, and problem gambling. Psychological review 2007;114(3):784.

[32] Myers KM, Ressler KJ, Davis M. Different mechanisms of fear extinction dependent on length of time since fear acquisition. Learning & memory 2006;13(2):216–223.

[33] Kalman RE. A new approach to linear filtering and prediction problems 1960;.

[34] Gershman SJ. A unifying probabilistic view of associative learning. PLoS computational biology 2015;11(11):e1004567.

[35] Eldar E, Rutledge RB, Dolan RJ, Niv Y. Mood as representation of momentum. Trends in cognitive sciences 2016;20(1):15–24.

[36] Stickgold R. Sleep-dependent memory consolidation. Nature 2005;437(7063):1272–1278.

[37] Smolen P, Zhang Y, Byrne JH. The right time to learn: mechanisms and optimization of spaced learning. Nature Reviews Neuroscience 2016;17(2):77.

[38] Ebbinghaus H. Memory: A contribution to experimental psychology. Oxford, England: Dover; 1964.

[39] Lattal KM. Trial and intertrial durations in Pavlovian conditioning: issues of learning and performance. Journal of Experimental Psychology: Animal Behavior Processes 1999;25(4):433.

[40] Commins S, Cunningham L, Harvey D, Walsh D. Massed but not spaced training impairs spatial memory. Behavioural brain research 2003;139(1-2):215–223.

[41] Cepeda NJ, Pashler H, Vul E, Wixted JT, Rohrer D. Distributed practice in verbal recall tasks: A review and quantitative synthesis. Psychological bulletin 2006;132(3):354.

[42] Urcelay GP, Wheeler DS, Miller RR. Spacing extinction trials alleviates renewal and spontaneous recovery. Learning & Behavior 2009;37(1):60–73.

[43] Luu L, Stocker AA. Post-decision biases reveal a self-consistency principle in perceptual inference. Elife 2018;7:e33334.

[44] Osan R, Tort AB, Amaral OB. A mismatch-based model for memory reconsolidation and extinction in attractor networks. PloS one 2011;6(8):e23113.

[45] Bouton ME, Woods AM, Pineño O. Occasional reinforced trials during extinction can slow the rate of rapid reacquisition. Learning and Motivation 2004;35(4):371–390.

[46] van den Akker K, Havermans RC, Jansen A. Effects of occasional reinforced trials during extinction on the reacquisition of conditioned responses to food cues. Journal of Behavior Therapy and Experimental Psychiatry 2015;48:50–58.

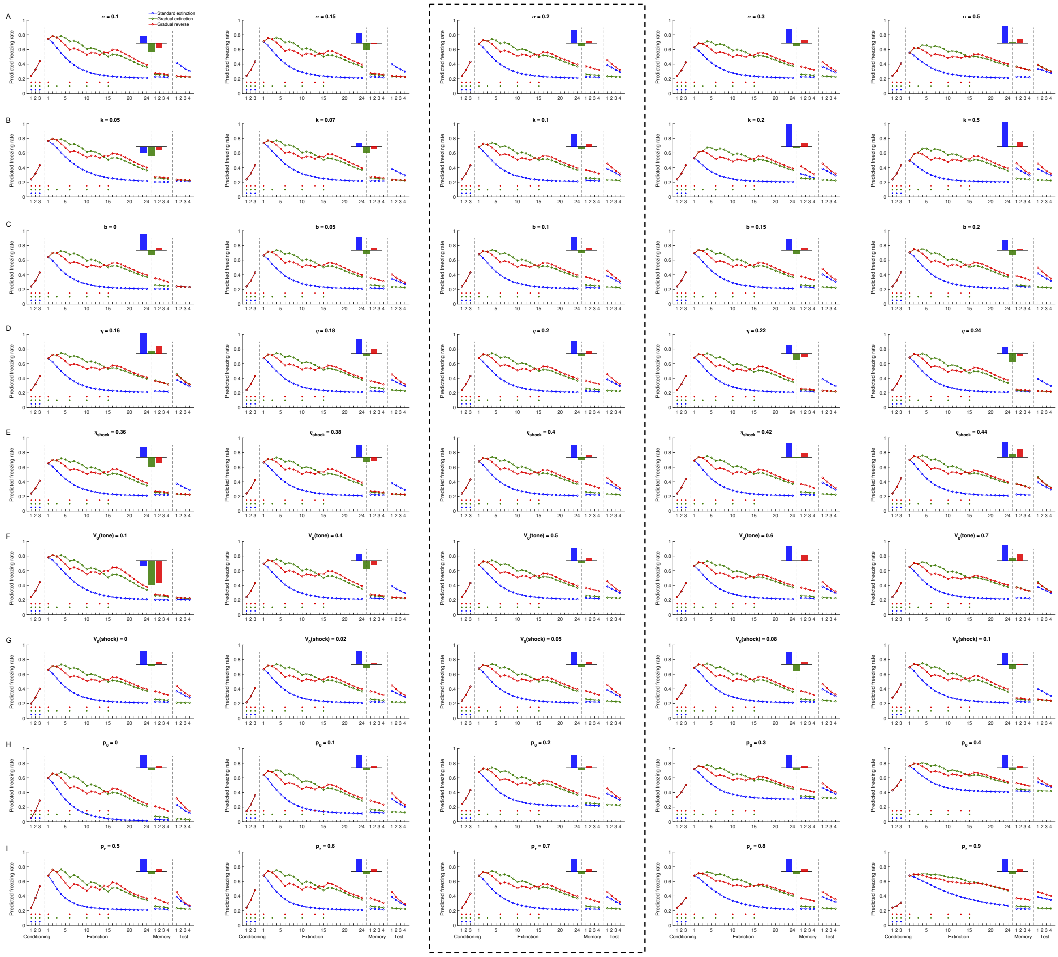
[47] Baldi E, Lorenzini CA, Bucherelli C. Footshock intensity and generalization in contextual and auditory-cued fear conditioning in the rat. Neurobiology of learning and memory 2004;81(3):162–166.

[48] Le Bars D, Gozariu M, Cadden SW. Animal models of nociception. Pharmacological reviews 2001;53(4):597–652.

[49] Gershman SJ, Hartley CA. Individual differences in learning predict the return of fear. Learning & behavior 2015;43(3):243–250.

[50] Gallistel CR, Fairhurst S, Balsam P. The learning curve: implications of a quantitative analysis. Proceedings of the National Academy of Sciences 2004;101(36):13124–13131.

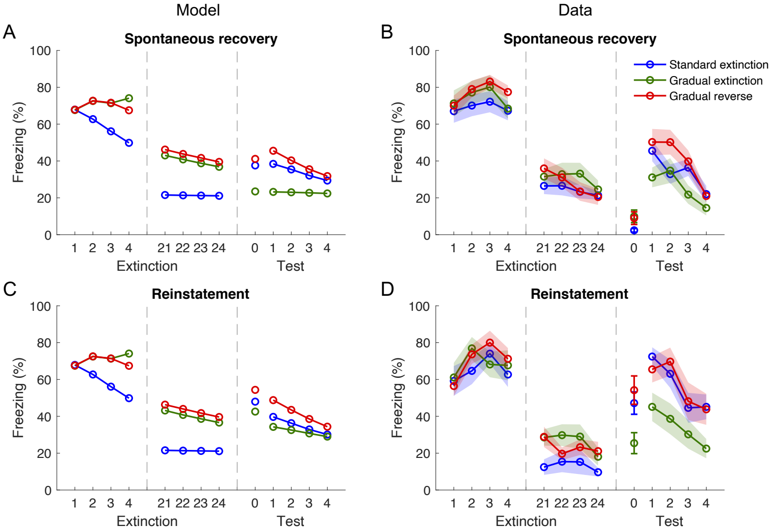
[51] Daw N, Courville A. The pigeon as particle filter. Advances in neural information processing systems 2008;20:369–376.



SUPPLEMENTARY FIGURE 1 Effects of different parameter values on the predicted pattern of results. We simulated the model by setting the values of all but one parameter as specified in Methods, and varying one parameter at a time. Each row of plots represents change of one parameter (as noted below and in sub-figure titles). For simplicity, we show only the spontaneous recovery experiment; results are consistent for the reinstatement experiment. Within each sub-figure, the main graph shows the predicted freezing rate over the course of the experiment (similar to Figure 4A), and the inset bar graph shows the log posterior ratio between two-cause and one-cause assignments on the last trial of extinction (ranging from -4 to 4, the same scale across all sub-figures), with positive and negative values showing an advantage of the two-cause assignment and the one-cause assignment, respectively; extinction procedures color-coded as in the main graph. The center column (same for all rows), in a dashed square, corresponds to the original parameter values as in the main model: main graphs in this column are the same as Figure 4A; insets show the same results as in Figure 3A, but only for the last extinction trial and in a bar graph. **(A)** $\alpha = [0.1, 0.15, 0.2, 0.3, 0.5]$ (from left to right, same below). α is the concentration parameter for the distance-dependent Chinese restaurant process (ddCRP) prior, which controls the prior probability of generating a new latent cause. In the current experiments, α determines the *a priori* probability ratio between forming two distinct causes and one single latent cause, shifting the curves in Figure 3A up (larger α's) or down (smaller α's), and therefore changing the final posterior ratio (as shown in inset). With smaller α's (e.g., $\alpha = 0.1, 0.15$), animals form only one latent cause in the gradual reverse condition, and therefore show no return of fear in that condition; large α values (e.g., $\alpha = 0.5$) lead to the two-cause beliefs for all extinction procedures, and therefore similar return of fear across all conditions.

(continues on next page)

SUPPLEMENTARY FIGURE 1 *(continued from previous page)*
(B) $k = [0.05, 0.07, 0.1, 0.2, 0.5]$. k is the decay parameter in the temporal kernel in the distance-dependent Chinese restaurant process prior, with a larger k corresponding to a smaller reactivation probability for an old latent cause. As a result, smaller k's (e.g., k = 0.05) make one-cause assignment more likely across all procedures, and therefore result in no return of fear. As k increases, two-cause beliefs first obtains dominance in standard extinction ($k = 0.07$), then in gradual reverse ($k = 0.1, 0.2, 0.5$), and eventually in gradual extinction if k is large enough (not shown). **(C)** $b = [0, 0.05, 0.1, 0.15, 0.2]$. b is the baseline probability for old causes in the distance-dependent Chinese restaurant process. Model simulation results are relatively insensitive to the value of b, unless very small (e.g., $b = 0$). In that case, all old causes become very unlikely after 30 days, so animals would infer a new latent cause during the spontaneous recovery test for all extinction procedures, resulting in no return of fear. **(D,E)** $\eta = [0.16, 0.18, 0.2, 0.22, 0.24]$, and $\eta_{\text{shock}} = [0.36, 0.38, 0.4, 0.42, 0.44]$, respectively. η and η_{shock} are learning rates in Rescorla-Wagner update rule (the presence of shocks has a higher learning rate of η_{shock}, compared to all other stimuli, including "tone", "no-tone" and "no-shock"). Learning rates determine the extent to which animals update their expectation after observing each stimulus. Specifically, in the current experiment, the comparison between η_{shock} and η is important in latent-cause inference. If animals learn about shocks with a relatively higher learning rate (corresponding to a higher value of η_{shock} or a lower value of η), they will be more likely to infer two latent causes (each with distinct shock frequencies, thanks to faster update), resulting in return of fear for all procedures; to the contrary, a lower η_{shock} or a higher η lead to slower updates and animals tend to infer one latent cause (i.e., no return of fear in the gradual reverse condition). **(F,G)** $V_0(\text{tone}) = [0.1, 0.4, 0.5, 0.6, 0.7]$, and $V_0(\text{shock}) = [0, 0.02, 0.05, 0.08, 0.1]$, respectively. $V_0(\text{tone})$ and $V_0(\text{shock})$ are the *a priori* frequency of tones and shocks for a newly inferred latent cause. We had set the parameter values according to the relative frequency of the two types of stimuli in animals' natural environments (tones are presumably more common than shocks). **(F)** If, however, the tone is assumed to be rarer for a new latent cause (i.e., lower values of $V_0(\text{tone})$), animals would be more likely to group all experiences into a single latent cause that is high in tone frequency. This is because all the trials are similar in having a tone, whereas a new cause would predict much lower tone frequency. As a result, the model predicts no return of fear for all procedures. To the contrary, if tones are assumed to be almost always present (e.g., $V_0(\text{tone}) \geq 0.7$), animals tend to infer new causes, which results in a two-cause belief and return of fear across all procedures. **(G)** Model simulation results are robust to small $V_0(\text{shock})$ values. If $V_0(\text{shock})$ is unrealistically high, however, animals would be unlikely to infer a second latent cause, as this new cause will not be able to account well for the "safe" experience. The dominance of the one-cause belief would then lead to no return of fear in the gradual reverse condition. **(H)** $p_0 = [0, 0.1, 0.2, 0.3, 0.4]$. p_0 is the baseline freezing rate after animals become aware of shocks. It does not affect latent-cause inference (i.e., the inset figures), but simply shifts the freezing rate curves up or down, in accordance with shifted baseline freezing seen empirically. **(I)** $p_r = [0.5, 0.6, 0.7, 0.8, 0.9]$. p_r is the perseveration parameter that controls the probability that animals will repeat the action from the previous trial. p_r does not affect latent-cause inference, but rather determines the "smoothness" of the behavior curves.



SUPPLEMENTARY FIGURE 2 Freezing rate comparison between model predictions (A,C) and empirical results (B,D). Shown are the first and last four trials of extinction, the beginning of the test session (before the first test trial; marked as trial 0), and the four test trials. Panels B and D are reproduced from [7]. Error bars and shaded areas in B,D indicate 1 s.e.m. across animals.