Selective maintenance of negative memories as a mechanism of spontaneous recovery of fear after extinction

Isabel M. Berwian[1,2,*], Sashank Pisupati[3], Jamie Chiu[2], Yongjing Ren[4] & Yael Niv[1,2]

[1]Princeton Neuroscience Institute & [2]Department of Psychology, Princeton University, USA
[3]Limbic Ltd, UK; [4]University Behavioral Health Care, Rutgers University, USA
[*]Corresponding author: Isabel Berwian (iberwian@princeton.edu)

Abstract

Spontaneous recovery of fear after extinction is a well-established behavioral phenomenon. Different theories in psychology account for spontaneous recovery by proposing that it may result from temporal weighting, reduced processing of stimuli over time, enhanced salience of adverse events or return of the acquisition context. We propose a novel mechanism of spontaneous recovery: selective maintenance of adverse events, and ground this mechanism in a computational model of latent cause inference. To investigate the proposed mechanism, we collected behavioral data with an aversive conditioning and extinction task (N=280) and fit the data with computational models formalizing our and others' theories. Quantitative and qualitative model comparisons indicated that selective maintenance of adverse events accounts for spontaneous recovery better than alternative theories. As spontaneous recovery of fear after extinction can serve as a model of relapse after exposure therapy, we use this mechanistic understanding of spontaneous recovery to propose and simulate the effect of add-on interventions to prevent relapse after exposure therapy.

Keywords: Computational Psychiatry; Spontaneous recovery; Latent cause inference; Exposure therapy

Introduction

Exposure therapy is the most effective treatment for anxiety disorders (Parker et al., 2018). The critical ingredient of exposure therapy is that the individual confront the feared situation or stimulus, but without the expected negative outcome. This "extinction" procedure presumably reduces the expectation of a negative outcome in the future and, thus, fear (Craske et al., 2014). However, fear often returns with time and clients may relapse (Craske & Mystkowski, 2006). Understanding why fear returns can help design interventions to prevent this from happening. Here, we present a novel mechanism of return of fear after extinction training and use it to explain how three modifications to exposure therapy may prevent relapse.

Fear conditioning and extinction paradigms in animals and humans inspired the development of exposure therapy and are widely used to study the mechanisms underlying its effectiveness. In such paradigms, the subject is typically first exposed to a neutral stimulus (conditional stimulus; CS) followed by an aversive stimulus (unconditional stimulus; US) during an acquisition phase. In the extinction phase, they are exposed to the CS without the US. After some time, the return of fear (i.e. spontaneous recovery) is assessed by presenting the CS alone in a test phase. Empirical evidence from such experiments indicate that spontaneous recovery is time-dependent, increasing with time since extinction (Rescorla, 2004).

Most researchers agree that spontaneous recovery requires learning of two competing associations (but see Paskewitz et al., 2022), one associating the CS with the US, which underlies acquisition of fear, and one that associates the CS with no adverse outcome, driving extinction of fear (Bouton, 1993; Gershman et al., 2013; Craske et al., 2008). In the test phase, the two associations compete. For spontaneous recovery to occur, the CS-US association must be stronger than the CS-noUS association. The most prominent existing theories of spontaneous recovery assume a weak CS-noUS association due to reduction of processing of the CS with more exposures (Pavlov, 1927), return of the acquisition context (Bouton, 1993) or temporal weighting (Devenport, 1998).

Here, we propose selective maintenance of negative memories (e.g., through conscious and subconscious replay) as a novel mechanism of spontaneous recovery. This idea is inspired by widely established behavioral evidence that emotionally valenced experiences are preferentially remembered over time (Dalgleish & Hitchcock, 2023; Rouhani et al., 2023), and neural evidence for replay as preventing forgetting (Wimmer et al., 2023). Selective replay of memories of adverse events might maintain them in the light of other memory decay processes and give them a competitive advantage in later retrieval.

To quantitatively compare ours and other theories of spontaneous recovery, we apply computational modeling to data we collected on an online aversive conditioning and extinction task. To model the creation of multiple competing associations, we use the latent cause framework (Gershman et al., 2010). This normative Bayesian framework proposes that individuals allocate observations (i.e., combinations of stimuli) to different latent causes based on their similarity to past experiences. The model quantifies the probability that events that are dissimilar from previous experience will lead to inference of a new latent cause (analogous to a new association; though associations in the model are between a latent cause and the observations it generates, not between the observations themselves; see Fig. 2). For instance, observations of CSs followed by USs may be assigned to one latent cause, whereas observations of (even those same) CSs without USs may be assigned to a different latent cause.

Using trial-by-trial model fitting, we compare different models, each formalizing a different mechanistic hypothesis. Fit to each participant separately, model parameters can cap-

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

ture individual differences in spontaneous recovery. Understanding the mechanism that gives rise to spontaneous recovery may therefore help tailor interventions to prevent return of fear based on each individual's learning parameters.

Method

Behavioral Task

We designed an online-administered aversive conditioning and extinction task (Fig. 1). The task was inspired by fear acquisition and extinction paradigms widely used in human and animal research, which are thought to mimic learning in exposure therapy. On each trial, participants saw one of two stimuli (CS+ or CS-), and had to press the space bar for the trial's outcome (an aversive but tolerable auditory scream US, or nothing). The task was completely Pavlovian; key pressing allowed reaction-time measurements and ensured continued attention to CSs. Every three trials, on average, participants rated how likely they expected each CS to be followed by a scream on a scale of 0-100% (29 ratings for each CS). In the acquisition phase (26 trials) 50% of 16 CS+ trials were followed by a US (total: 8 screams), and the CS- appeared on 12 trials. After a 3-5 minute break (in which participants filled out questionnaires) the extinction phase began (30 trials; 18 CS+, 12 CS-, no US). Next, participants completed a separate task for ∼15 minutes, followed by a spontaneous recovery test (16 trials CS+ or CS-; no US), and then a relearning phase (10 CS+ with 4 USs; 6 CS-). Trial order was fixed for all participants. To ensure remote participants heard the US loudly enough, six auditory attention checks were spread throughout the task, and at the end of the task they were asked to report if they had changed the volume during the experiment.

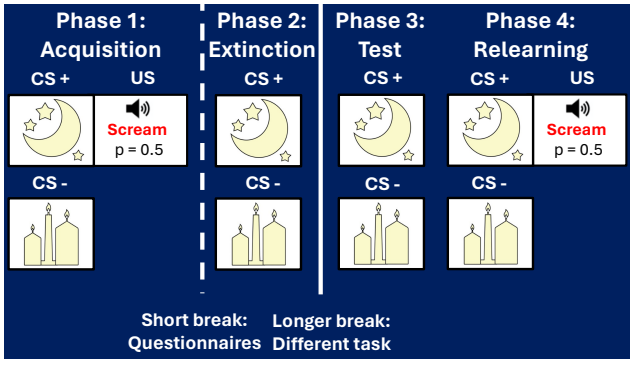

Figure 1: **Task design**. On each trial, participants observed a moon (CS+) or candle (CS-). In the acquisition and relearning phase, the moon was followed by an aversive scream (US) on half the trials. Otherwise, there were no USs.

Procedure and participants

This study was approved by the Institutional Review Board of Princeton University, and all participants provided written informed consent. Over 700 participants were recruited from Prolific to complete several online behavioral tasks and mental health questionnaires in spring 2023. Here, we analyze the first N=280 participants with complete and accurate datasets of the aversive learning task. Participants were compensated for their time (∼ 50 minutes at $13/hr). They had to reside in the United States, Canada, Australia, or New Zealand, be fluent in English, and have headphones. Data from 38 participants who failed more than one audio attention check and 29 who indicated they changed volume were excluded. The final dataset contained N=213 subjects.

Data visualization. To illustrate individual differences in behavior that might reflect differences in underlying mechanisms, we separately plotted four "groups" of participants. Behavioral cut-offs for these were decided *a priori* based on pilot data. First, participants who did not show differential (>10%) expectations for the CS+ and CS- by the end of acquisition were termed 'Generalizers' (N=9). Remaining participants were then divided into groups based on the spontaneous recovery test: N=127 participants who increased their US expectation for the CS+ by more than 10% compared to the end of extinction, but not so for the CS-, were termed the 'spontaneous recovery' ('SR') group. N=56 participants who did not increase expectation for either stimulus by more than 10% were termed the 'No SR' group, and N=18 who increased expectations for both stimuli were termed the 'Return to Prior' group. Three participants did not meet any of these criteria and were not included in subgroup analyses.

Generative models

We formalized learning of the latent structure of the task using Bayesian nonparametric models of latent-cause inference (Gershman et al., 2015; Gershman & Blei, 2012). In the latent-cause framework, we attribute observations (here, trials) to one of an unlimited number of latent causes $L_t \in [1,...,j]$, each with its own unique set of parameters $\Phi_j = \{\phi_{j,i}\}$ that determine probabilities of observing each of $i \in [1:4]$ features. We set i=1 to indicate the CS-, i=2 the CS+, i=3 a 'break stimulus' (for the time between acquisition and extinction and between extinction and test) and i=4 the US. Thus, each latent cause embodies a different association – different probabilities of observing the CSs and US.

Prior over latent causes. An infinite-capacity prior over latent causes can flexibly add new causes when they are necessary to explain dissimilar observations. We used a Chinese Restaurant Process (CRP) prior (Aldous et al., 1985) that generates observations on trial $t+1$ by first selecting a latent cause L_{t+1} according to:

$$p(L_{t+1} = j|\mathbf{L}_{1:t}) = \begin{cases} \frac{N_{j,t}}{\Sigma_{m=1}^{t} N_{m,t}+\alpha} & \text{if } j \text{ is an old latent cause} \\ \frac{\alpha}{\Sigma_{m=1}^{t} N_{m,t}+\alpha} & \text{if } j \text{ is a new latent cause} \end{cases}$$

where $N_{j,t}$ is the number of observations generated by latent cause j up to trial t (see more below) and $\alpha \geq 0$ is a parameter that influences the probability of new latent causes.

Prior over observations. Next, observations $\mathbf{O}_{t+1}$ are generated by independent Bernoulli processes with probability

$\phi_{j,i}$ for each feature i, conditioned on the latent cause drawn for the trial, $p(O_{i,t+1}|L_{t+1}=j)=\phi_{j,i}(t+1)$. When a new latent cause j is initialized, initial probabilities $\phi_{j,i}$ are drawn from Beta priors with two parameters $a_{obs}>0$ and $b_{obs}>0$. Here, given no bias in the data (i.e., 50% expectation of US on the first trial), we assume symmetric Beta priors ($a_{obs}=b_{obs}$). Larger a_{obs} lead to $\phi_{j,i}$ that are more stochastic (e.g., around 0.5; Fig. 2A) while smaller a_{obs} lead to $\phi_{j,i}$ that are more deterministic (closer to 0 or 1; Fig. 2C). Later, $\phi_{j,i}(t)$ are updated according to observed events (see eq. 6 below).

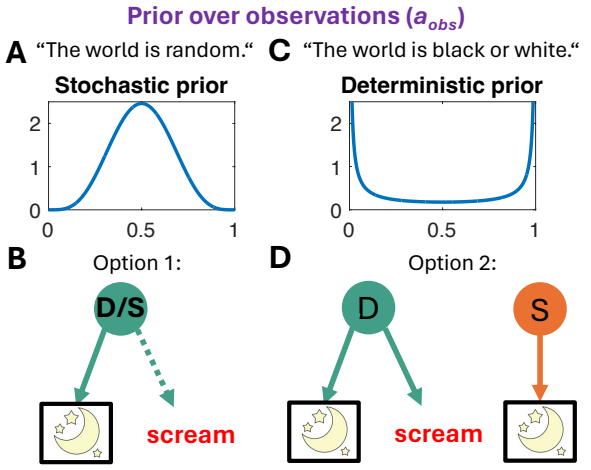

Figure 2: **Model illustration**. Consider a participant who has already assigned the moon (CS+) and the scream (US) to latent cause "D" (for dangerous). **A)** If they have a stochastic prior ($a_{obs}>=1$), when observing a moon without a US **B)** they will be more likely to infer this was also generated by D, and update the probability of the US given that latent cause (now D/S – partly dangerous and partly safe;). This is because their prior allows for observations to be probabilistic. **C)** In contrast, if they have a deterministic prior ($a_{obs}<1$), the absence of the US will likely lead to **D)** inference of a new safe ("S") latent cause that is only associated with the CS+. The dangerous latent cause will then remain unchanged, despite the safe experience.

The basic model. Here, $N_{j,t}$ is the number of trials caused by latent cause j up to trial t. Since there is nonzero probability for each of the previous latent causes to have generated the current trial, in practice $N_{j,t}$ sums the posterior probability of latent cause j on each trial: $N_{j,t}=\Sigma_{t'\leq t}P(L_{t'}=j|\mathbf{O}_{1:t})$ where $\mathbf{O}_{1:t}$ are the observations up to trial t. We fit two versions of this model, a "basic α model" with α as a free parameter and a "basic prior model" with a_{obs} as a free parameter.

Decay model. To account for decay of memory over time, following Blei and Frazier (2011), we modified the basic model to decay the counts $N_{j,t}$ with a rate determined by $0\leq\gamma\leq 1$ as follows: $N_{j,t}=\Sigma_{t'\leq t}P(L_{t'}=j|\mathbf{O}_{1:t})\gamma^{(t-t')}$. This model reduces to the basic model when $\gamma=1$.

Selective maintenance model. Inspired by evidence that humans remember negative events better (Rouhani et al., 2023), in this model we hypothesize that latent causes associated with aversive stimuli (like the scream US) are protected from decay[1], perhaps due to memory replay (Wimmer et al., 2023). We model this by reducing the decay rate (i.e., increasing the effective γ towards 1) according to a parameter $0\leq\omega\leq 1$, scaled by the estimated probability of the US given the latent cause:

$$N_{j,t}=\Sigma_{t'\leq t}P(L_{t'}=j|\mathbf{O}_{1:t})\cdot\prod_{k=t'+1}^{t}d_{j,k} \quad (1)$$

$$d_{j,k}=\gamma+(1-\gamma)\omega\cdot p(O_4|L_k=j). \quad (2)$$

Here, $p(O_4|L_k=j)$ is given by $\phi_{j,4}$, which is updated on each trial (see likelihood computation below, eq. 6). This model reduces to the decay model when $\omega=0$.

Temporal weighting model. To compare our hypothesis to the temporal weighting rule suggested to account for spontaneous recovery (Devenport, 1998), we modeled power law decay of counts with a rate determined by $\iota\geq 0$:

$$N_{j,t}=\Sigma_{t'\leq t}P(L_{t'}=j|\mathbf{O}_{1:t})*\frac{1}{(t+1-t')^{\iota}} \quad (3)$$

This model reduces to the basic model for $\iota=0$.

Salience model. Emotional events are more salient and capture more attention (see Dolcos et al. (2020), for a review), potentially enhancing memory of these events. We therefore implemented a model that encodes aversive experiences as $(1+\varepsilon)$ rather than 1 in the counts $N_{j,t}$,

$$N_{j,t}=\Sigma_{t'\leq t}P(L_{t'}=j|\mathbf{O}_{1:t})*(1+O_{4,t'}\cdot\varepsilon) \quad (4)$$

with $\varepsilon\geq 0$. Note that $O_{4,t}$ is 1 on trials with a US, and 0 otherwise. This model reduces to the basic model for $\varepsilon=0$.

Processing loss model To model the idea that repeated exposure to a stimulus may reduce its processing due to habituation or neural fatigue (Pavlov, 1927), we set the count $N_{j,t}$ according to a decreasing logistic function of how often that CS (O_1 or O_2, one observed per trial) had been seen so far:

$$N_{j,t}=\Sigma_{t'\leq t}P(L_{t'}=j|\mathbf{O}_{1:t})*(d_{1,t'}O_{1,t'}+d_{2,t'}O_{2,t'}) \quad (5)$$

where $d_{i,t}=\frac{1}{1+e^{-\lambda(\nu+\Sigma_{t'\leq t}O_{i,t'})}}$ and $\lambda\leq 0$ and $\nu\geq 0$. This model reduces to the standard model as $\nu\to\infty$.

Likelihood of observations. For all models, we assumed that the Bernoulli likelihoods are updated after every trial by summing the occurrences of features of all previous trials weighted by the probability of the latent cause (with a_{obs} as the prior "occurrence") and normalizing:

$$\phi_{j,i}(t)=\frac{a_{obs}+\sum_{t'\leq t}O_{i,t'}\cdot P(L_{t'}=j|\mathbf{O}_{1:t'})}{2a_{obs}+\sum_{t'\leq t}P(L_{t'}=j|\mathbf{O}_{1:t'})}. \quad (6)$$

[1]Note that the current data cannot differentiate between selective maintenance of aversive events and faster decay rates of inhibitory associations as proposed by Paskewitz et al. (2022).

Model fitting

We fit the free parameters of each of the models above (Table 1) to the behavioral data of each individual participant separately. As these nonparametric models are analytically intractable, we used a simulation-based approach to generate predictions of the occurrence of the US (the scream) and estimated parameters by maximizing the likelihood of the participant's ratings assuming the response error was normally distributed around the model prediction with a fixed $\sigma = 0.15$ (assuming instead that the response error followed a beta distribution led to the same pattern of results). We used the Bayesian Adaptive Direct Search (BADS; Acerbi & Ma, 2017) algorithm for parameter search, with 12 restarts for each subject and model. To constrain the search space, parameter bounds were: $0 < \alpha \leq 10$, $0 < a_{obs} \leq 10$, $0 < \gamma \leq 1$, $0 < \omega \leq 1$, $0 < \iota \leq 5$, $0 < \varepsilon \leq 10$, $-10 \leq \lambda < 0$, $0 < \nu \leq 100$.

Table 1: Model parameters.

Model name	Free Θ	Fixed Θ
Basic α model	α	$a_{obs} = 1$
Basic prior model	a_{obs}	$\alpha = 0.03$
Decay model	a_{obs}, γ	$\alpha = 0.03$
Selective maintenance model	a_{obs}, γ, ω	$\alpha = 0.03$
Temporal weighting model	a_{obs}, ι	$\alpha = 0.03$
Salience model	a_{obs}, ε	$\alpha = 0.03$
Processing loss model	a_{obs}, λ, ν	$\alpha = 0.03$

Inference simulations. For each model, to infer a distribution over latent causes given observations, we approximated Bayesian latent cause inference using particle filtering (sequential importance resampling of 1000 particles). On each trial, after observing the CS, the algorithm inferred the posterior probability over possible partitions of trials into latent causes based on all previous observations and used this to generate a prediction of the probability of the US:

$$p(O_{4,t} = 1) = \Sigma p(O_{4,t} = 1 | L_{1:t}) * p(L_{1:t} | O_{1:3,1:t}, O_{4,1:t-1}, \Theta, M)$$

where the sum is over all possible latent cause assignments approximated by averaging across particles.

After observing whether a scream US indeed occurred on that trial, the posterior distribution over latent causes was recalculated and used to update the observation probabilities for each latent cause $\phi_{j,i}$. Throughout, the model observed the same stimuli as participants. To simulate breaks, we presented the model with a dummy stimulus O_3 for 9 trials between acquisition and extinction (short break) and 34 trials before the spontaneous recovery test phase (long break).

Model comparison. To identify the winning model, we compared median BICs between non-nested models and used likelihood ratio tests for nested models. We used nonparametric significance tests when data were not normally distributed.

Results

Behavioral results

We replicated our previous results (Pisupati et al., 2023) showing that on average, in acquisition, participants learned to differentiate between the CS+ and the CS- and correctly predicted that a scream was likely to follow the CS+ (Fig. 3A, orange), but not the CS- (Fig. 3A, light blue). In extinction, they decreased their expectations of the scream, however, many participants showed increased scream expectancy for the CS+ in the spontaneous recovery test. Finally, participants quickly relearned that the CS+ was followed by the scream in the relearning phase.

When plotting groups of participants based on their behavioral responses in the acquisition and test phases (see Data Visualization in Method; Fig. 3C), the SR group (participants who showed increased expectation of the scream for the CS+ but not CS- in the test phase) also showed faster relearning compared to the No SR group (participants who did not increase expectations of the scream at test). The Return to Prior group, who expected a scream for both CS+ and CS- in the test phase, increased their expectancy ratings to 50% (hence, to the prior at the start of the experiment). Finally, the Generalizers group did differentiate the CS+ and CS- in the relearning phase suggesting they were not inattentive. Note also that these participants also evidenced learning as they reduced US expectations during extinction.

Model fitting results

Model comparison. Wilcoxon sign-rank tests showed both the decay and selective maintenance models had significantly lower median BICs than the next best model, i.e., the temporal weighting model ($Z = -7.817$, $p < 0.001$ and $Z = -4.8846$, $p < 0.001$, respectively, Fig. 4A). Likelihood ratio tests for the nested decay, selective maintenance, and basic prior models favored the decay model over the basic prior model ($\chi^2(213) = 3083$, $p < 0.001$) and the selective maintenance model over the decay model ($\chi^2(213) = 2102$, $p < 0.001$). Likelihood ratio tests at the level of individual subjects indicated that the basic prior model fit 98 of 213 subjects better than the decay model and 51 of 213 subjects better than the selective maintenance model. Furthermore, the decay model fit 151 of 213 subjects better than the selective maintenance model, indicating robust individual differences ($p < 0.05$ for all comparisons). Of the 62 subjects better fit by the selective maintenance model, 45 were in the SR group and only 9 in the no SR group, which was significantly different from chance ($\chi^2(183) = 7.004$, $p < 0.008$). As evident from Fig. 4B, the selective maintenance model was the only model that captured all features of the behavior, specifically including the amount of spontaneous recovery observed in the data. It also captured all behavioral features of the different subgroups, as shown in Fig. 3C (red and dark blue curves).

Parameter comparison. To uncover the mechanisms that give rise to spontaneous recovery and other behavioral features of each of the subgroups, we compared the estimated

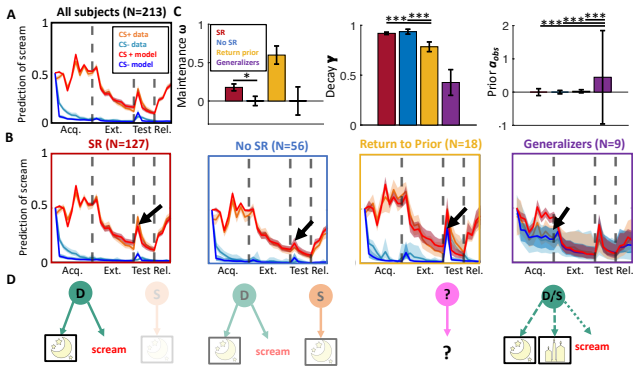

Figure 3: **Behavioral and modeling results**. **A)** Mean expectancy ratings for the scream after the CS+ (orange) and the CS- (light blue) over time for the four phases of the task (separated by dashed vertical lines; Acq.: acquisition, Ext.: extinction, Test: spontaneous recovery test and Rel.: relearning) across all subjects. Red and dark blue curves are model predictions from the selective maintenance model averaged across participants. Shading: 95% bootstrapped confidence intervals. **B)** Same as A but visualized separately for participants who showed spontaneous recovery for CS+ only (SR group), those who did not (No SR group), those who showed high expectations for both stimuli at test (Return to Prior group) the Generalizers group, who did not differentiate the CS+ from CS- in acquisition. **C)** Median parameter estimates for the selective maintenance model for participants in each of the four groups. *: $p < 0.05$, ***: $p < 0.001$. **D)** Illustration of the inferred latent causes that give rise to the prediction indicated with a black arrow in each group in B.

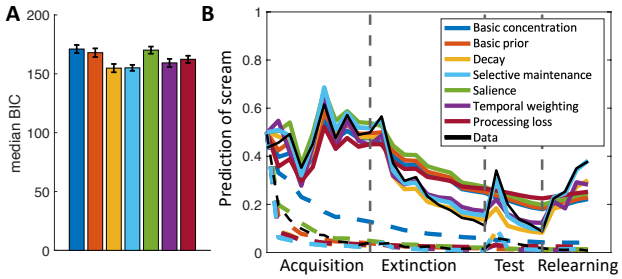

Figure 4: **Model comparison**. **A)** Median BIC of each model for all subjects. Lower BIC scores indicate better model fits. The decay (yellow) and selective maintenance (light blue) models fit the data best. As they are nested, we further compared these models using likelihood ratio tests (main text). See model-color mappings in the legend of B. **B)** Expectation of the scream for the moon (CS+, solid lines) and the candles (CS-, dashed lines). Black: empirical data averaged over all subjects. Colors: simulations using best parameter estimates for each model. Vertical dashed lines indicate the end of a phase in the task. The selective maintenance model (light blue), was the only model that quantitatively captured both the spontaneous recovery at the beginning of the test phase and rapid reacquisition in the relearning phase.

parameters of the selective maintenance model across subgroups. Before analyzing this dataset, we had formulated three independent hypotheses based on a previous pilot behavioral datasets: 1) The SR group will show selective maintenance of latent causes associated with the scream US, leading to increased probability of these latent causes in the spontaneous recovery test. 2) The Return to Prior group will show rapid decay of latent causes, which will lead to the creation of a new latent cause after the break, in the test phase. 3) The Generalizers will have a stochastic prior over observations, thereby encouraging inference of a single latent cause that accounts for both CS+ and CS- trials.

Using Wilcoxon ranksum tests, we found evidence for all three *a priori* hypotheses. Estimates of the selective maintenance parameter ω were higher in the SR group than in the No SR group ($Z = -2.0669$, $p = 0.039$; Fig. 3B, left). The Return to Prior group had significantly lower decay parameter estimates (indicating faster decay) than the SR and No SR groups ($Z = -3.798$, $p = 0.001$, and $Z = -3.584$, $p = 0.001$, respectively; Fig. 3B, middle). Finally, as seen in Fig. 3B, right, the α_{obs} parameter estimates confirmed that the Generalizers group had more stochastic priors than the SR group ($Z = 4.0881$, $p < 0.001$), the No SR group ($Z = 3.979$, $p < 0.001$) and the Return to Prior group ($Z = 3.214$, $p = 0.001$). Note, however, that comparisons with the Generalizers group must be interpreted with great caution due to the small size of that group ($N = 9$). This pattern of results was also seen in other models, to the extent that each model included the relevant free parameters.

To ensure parameters were reliably recoverable from the data, we simulated data from the selective maintenance model using the parameters estimated for participants, and then fit the model to the simulated data. The correlations for the ground truth and recovered parameters were: $\omega : r = 0.69$, $\gamma : r = 0.98$, $a_{obs} : r = 1$. As ω was not independent of γ, we re-parameterized the model as $\omega' = (1-\gamma) \cdot \omega$. Ground truth and recovered parameter estimates of ω' were highly correlated ($r = 0.96$). Thus, with all $r's > 0.95$, we consider the parameters of the model highly reliable.

Examining the latent cause structure inferred by the selective maintenance model for each group of participants, we found further support for our hypotheses. In the SR group, latent causes strongly associated with the scream US gained relative strength throughout the task compared to other latent causes (e.g., those predicting the CS+ but not the US; Fig. 3D, left). This accounted for spontaneous recovery of US expectations in the test phase. Conversely, latent causes in the Return to Prior group decayed sufficiently over the long break such as to lead to inference of new latent causes in the test phase (Fig. 3D, pink). Finally, the Generalizers group inferred one latent cause that generated all observed features (Fig. 3D, right). Note that model parameters were fit to all trials without special weighting of the spontaneous recovery test; indeed, other patterns in the data were also explained by the parameters (not detailed here).

Discussion

We introduced selective maintenance of adverse events as a novel mechanism of spontaneous recovery and developed a normative Bayesian model that incorporates this mechanism. This model qualitatively captured all behavioral signatures in a large empirical dataset from an aversive conditioning and extinction task. It also explained the data better than models incorporating alternative mechanisms such as temporal weighting, increased salience of adverse events, and decreased processing of familiar stimuli over time. In line with our *a priori* hypotheses, we provided computational evidence that selective maintenance of adverse events can explain spontaneous recovery of CS+-specific US expectations at test; that recovery of US expectations for both CS+ and CS- can be explained by decay of all latent causes; and that generalization across CSs can be captured by partitioning experience into fewer latent causes. We previously showed that participants showing such generalization across CSs also show wider generalization to new stimuli (Aitsahalia, 2022), similar to people with anxiety disorders (Cooper et al., 2022).

Our results suggest that two conditions need to be jointly fulfilled for spontaneous recovery to occur: Participants must 1) partition their experiences of neutral and adverse events into different latent causes (in our model, this is driven by deterministic prior beliefs), and 2) selectively maintain the causes that led to adverse events. This proposed mechanism is neurobiologically plausible and evolutionarily adaptive as it can ensure that aversive learning is long-lived. In line with evidence that replay protects memories from being forgotten (Wimmer et al., 2023), we suggest that (some) people may preferentially replay memories (i.e., latent causes) of aversive events, thus maintaining them in the face of decay processes.

Based on the idea that extinction and spontaneous recovery are models of exposure therapy and relapse in anxiety disorders, we propose this understanding can be used to potentially improve exposure therapy interventions. If the two conditions above are needed, preventing one of them should be sufficient to prevent relapse. For instance, cognitive restructuring treatment that targets "black-and-white thinking" may change deterministic beliefs (i.e., increase α_{obs} to allow more stochastic observations) and help prevent inference of a new latent cause during exposure therapy (Smith et al., 2021). Longer inter-session intervals (Laborda et al., 2011) or the introduction of a retrieval cue (e.g., a wristband) that can serve as a reminder of the exposure experience (Craske et al., 2014) may help strengthen the memory of neutral latent causes and thus reduce the effect of selective maintenance of latent causes associated with adverse events. Fig. 5 shows simulations of these interventions, and their capacity to prevent spontaneous recovery, at least in our model.

We compared our hypothesis to a set of theories that have been put forward to explain spontaneous recovery[2], implementing all alternatives within the latent cause framework.

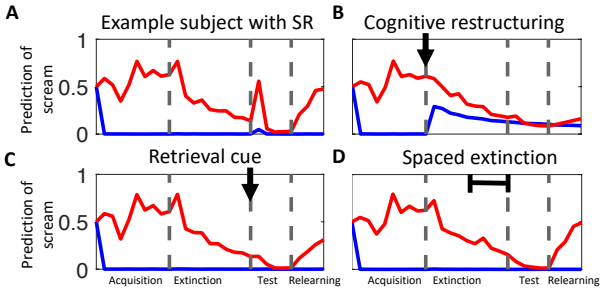

Figure 5: **Simulated effects of different interventions**. **A**) Simulated data with parameter estimates from a participant who showed spontaneous recovery. For this participant, we simulated the effect of **B**) cognitive restructuring by increasing a_{obs} by 3.5 to decrease "black-and-white thinking" before exposure (extinction); **C**) a retrieval cue after exposure (extinction), simulated by setting the decay rates for all latent causes to the same value from then forward, and **D**) spaced exposure simulated by moving some extinction trials from the second half of extinction into the break phase. Black symbols indicate when the intervention was applied. In all cases, the manipulation decreased spontaneous recovery.

The most prominent theory of spontaneous recovery proposes that we infer separate temporal contexts in acquisition and extinction (Bouton, 1993) with the return of the acquisition context leading to spontaneous recovery. Previously, we formalized a version of Bouton's temporal context hypothesis as a hierarchical Bayesian inference model to explain spontaneous recovery as resulting from temporal persistence of contexts and return of the acquisition context after context switches (Pisupati et al., 2023). This model, however, predicted less spontaneous recovery than observed in our empirical data and fit the data worse than the selective maintenance model (results not shown). Still, spontaneous recovery might be multi-determined (Rescorla, 2004), and apparent context changes or increased salience of adverse events combined with temporal weighting could also account for some of its behavioral signatures. That these mechanisms, when formalized so that they make quantitative predictions, do not account for the amount of spontaneous recovery observed in our and other datasets (Bouton, 1993; Quirk, 2002), suggests selective maintenance may nevertheless be at play.

Finally, we note that our model, with parameters fit to all trials without preferential weighting of the test phase, predicted spontaneous recovery for only approximately half the SR group. This suggests that a combination of mechanisms might explain different participants, and/or additional mechanisms we have not yet modeled. We look forward to uncovering these mechanisms in future work.

[2]One set of theories that we did not model proposes that extinction of responding is due to neural (Pavlov, 1927) or behavioral fatigue (Rescorla, 2004). Responding then recovers with time, as fatigue wears off. Such theories cannot account for extinction of expectations in our task, as responses indicating high or low expectation of a scream required the same amount of effort.

Acknowledgements

Isabel Berwian, Yongjing Ren and Yael Niv's work on "Precision psychiatry for treatment selection in depression" is supported by Wellcome Leap as part of the Multi-Channel Psych Program. We would like to thank Sebastian N. Quinard for building the auditory components of the task and Nathanial Daw for discussions on the analysis.

References

Acerbi, L., & Ma, W. J. (2017). Practical Bayesian optimization for model fitting with Bayesian adaptive direct search. *Advances in Neural Information Processing Systems*, *30*, 1834–1844.

Aitsahalia, I. (2022). *Controllability priors modulating over- and under-segmentation of latent causes in fear conditioning*. Princeton university senior theses, Princeton University.

Aldous, D. J., Ibragimov, I. A., Jacod, J., & Aldous, D. J. (1985). *Exchangeability and related topics*. Springer.

Blei, D. M., & Frazier, P. I. (2011). Distance dependent chinese restaurant processes. *Journal of Machine Learning Research*, *12*(8).

Bouton, M. E. (1993, July). Context, time, and memory retrieval in the interference paradigms of Pavlovian learning. *Psychological Bulletin*, *114*(1), 80–99. doi: 10.1037/0033-2909.114.1.80

Cooper, S. E., van Dis, E. A., Hagenaars, M. A., Krypotos, A.-M., Nemeroff, C. B., Lissek, S., ... Dunsmoor, J. E. (2022). A meta-analysis of conditioned fear generalization in anxiety-related disorders. *Neuropsychopharmacology*, *47*(9), 1652–1661.

Craske, M. G., Kircanski, K., Zelikowsky, M., Mystkowski, J., Chowdhury, N., & Baker, A. (2008, January). Optimizing inhibitory learning during exposure therapy. *Behaviour Research and Therapy*, *46*(1), 5–27. doi: 10.1016/j.brat.2007.10.003

Craske, M. G., & Mystkowski, J. L. (2006). Exposure Therapy and Extinction: Clinical Studies. In *Fear and learning: From basic processes to clinical implications* (pp. 217–233). Washington, DC, US: American Psychological Association. doi: 10.1037/11474-011

Craske, M. G., Treanor, M., Conway, C. C., Zbozinek, T., & Vervliet, B. (2014, July). Maximizing exposure therapy: an inhibitory learning approach. *Behaviour Research and Therapy*, *58*, 10–23. doi: 10.1016/j.brat.2014.04.006

Dalgleish, T., & Hitchcock, C. (2023, March). Transdiagnostic distortions in autobiographical memory recollection. *Nature Reviews Psychology*, *2*(3), 166–182. (Number: 3 Publisher: Nature Publishing Group) doi: 10.1038/s44159-023-00148-1

Devenport, L. D. (1998). Spontaneous recovery without interference: Why remembering is adaptive. *Animal Learning & Behavior*, *26*, 172–181.

Dolcos, F., Katsumi, Y., Moore, M., Berggren, N., de Gelder, B., Derakshan, N., ... others (2020). Neural correlates of emotion-attention interactions: From perception, learning, and memory to social cognition, individual differences, and training interventions. *Neuroscience & Biobehavioral Reviews*, *108*, 559–601.

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

Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning, and extinction. *Psychological Review*, *117*, 197–209. (Place: US Publisher: American Psychological Association) doi: 10.1037/a0017808

Gershman, S. J., 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*, 164. doi: 10.3389/fnbeh.2013.00164

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

Laborda, M. A., McConnell, B. L., & Miller, R. R. (2011, March). Behavioral Techniques to Reduce Relapse After Exposure Therapy. In T. R. Schachtman & S. S. Reilly (Eds.), *Associative Learning and Conditioning Theory: Human and Non-Human Applications* (p. 0). Oxford University Press. doi: 10.1093/acprof:oso/9780199735969.003.0025

Parker, Z. J., Waller, G., Duhne, P. G. S., & Dawson, J. (2018). The role of exposure in treatment of anxiety disorders: A meta-analysis. *International Journal of Psychology & Psychological Therapy*, *18*(1), 111–141. (Place: Spain Publisher: Asociación de Análisis del Comportamiento)

Paskewitz, S., Stoddard, J., & Jones, M. (2022). Explaining the return of fear with revised rescorla-wagner models. *Computational Psychiatry*, *6*(1).

Pavlov, I. (1927). *Conditioned reflexes* (Vol. 430). London: Oxford University Press.

Pisupati, S., Berwian, I. M., Chiu, J., Ren, Y., & Niv, Y. (2023, 22–25 Aug). Human inductive biases for aversive continual learning — a hierarchical bayesian nonparametric model. In S. Chandar, R. Pascanu, H. Sedghi, & D. Precup (Eds.), *Proceedings of the 2nd conference on lifelong learning agents* (Vol. 232, pp. 337–346). PMLR.

Quirk, G. J. (2002). Memory for extinction of conditioned fear is long-lasting and persists following spontaneous recovery. *Learning & Memory (Cold Spring Harbor, N.Y.)*, *9*(6), 402–407. doi: 10.1101/lm.49602

Rescorla, R. A. (2004). Spontaneous recovery. *Learning & Memory (Cold Spring Harbor, N.Y.)*, *11*(5), 501–509. doi: 10.1101/lm.77504

Rouhani, N., Niv, Y., Frank, M. J., & Schwabe, L. (2023, September). Multiple routes to enhanced memory for emotionally relevant events. *Trends in Cognitive Sciences*, *27*(9), 867–882. doi: 10.1016/j.tics.2023.06.006

Smith, R., Moutoussis, M., & Bilek, E. (2021, May). Simulating the computational mechanisms of cognitive and

behavioral psychotherapeutic interventions: insights from active inference. *Scientific Reports*, *11*(1), 10128. doi: 10.1038/s41598-021-89047-0

Wimmer, G. E., Liu, Y., McNamee, D. C., & Dolan, R. J. (2023, February). Distinct replay signatures for prospective decision-making and memory preservation. *Proceedings of the National Academy of Sciences*, *120*(6), e2205211120. (Publisher: Proceedings of the National Academy of Sciences)