Selective maintenance of aversive memories as a mechanism of spontaneous recovery of fear

Isabel M. Berwian*[1,2], Yongjing Ren[1], Sashank Pisupati[2], Jialing Ding[1], Seohyun Moon[1], Jamie C. Chiu[2], Deepta Chandrasekhar[1], and Yael Niv[1,2]

[1]Princeton Neuroscience Institute, Princeton University, USA
[2]Department of Psychology, Princeton University, USA
[3]Atla AI Ltd, London, UK

Abstract

Return of fear after exposure poses a significant challenge for treatment of anxiety disorders. In this study, we used computational modeling to test competing mechanisms underlying spontaneous recovery of fear over time. We fit computational models of a novel theory of spontaneous recovery—selective maintenance of aversive memories—to behavior from a fear conditioning and extinction task (N=316), and showed that they uniquely captured spontaneous recovery and quantitatively outperformed alternative models embodying theories from the literature. The results were supported across multiple datasets, including a preregistered replication (N=355) and a sample with mental health symptoms (N=520). The selective maintenance modeling framework additionally offers mechanistic insights into overgeneralization and the development of anxiety. Indeed, in the symptomatic sample we found that symptoms of generalized anxiety disorder correlated with estimates of overgeneralization in the model. Through simulations, we further demonstrated that insights from our model can explain how targeted interventions such as retrieval cues and cognitive interventions can prevent the return of fear. These results highlight selective maintenance of aversive events in memory as a critical and testable target for improving anxiety treatments and preventing relapse.

Introduction

Exposure therapy is the most effective therapy for anxiety disorders (Parker, Waller, Duhne, & Dawson, 2018). During exposure, an individual confronts a feared stimulus

*Corresponding author: iberwian@princeton.edu or situation without experiencing the expected negative outcome. A common assumption is that learning from this experience leads to decreased prediction of the negative outcome, and thus a reduction of fear. However, fear often returns after some time, and up to 62% of individuals relapse (Craske & Mystkowski, 2006). To develop clinical methods that prevent relapse, it is critical that we understand why fear returns.

Exposure has been studied extensively with well established fear conditioning and extinction paradigms. In such paradigms, in the acquisition phase, a neutral stimulus, often called a CS+ (for conditional stimulus), is paired with an aversive stimulus (US; unconditional stimulus), leading to fear responses when confronted with the CS+. In the subsequent extinction phase, mimicking exposure therapy, the CS+ is presented without the US. This procedure leads to a reduction in fear of the CS+. Strikingly, after the passage of time and even without new exposure to the CS+ or US, fear of the CS+ often returns. This phenomenon of 'spontaneous recovery' is well established across species and seems to be time-dependent (Rescorla, 2004; Quirk, 2002; Robbins, 1990; Haberlandt, Hamsher, & Kennedy, 1978).

Fear can only return if the 'fear memory' (i.e., a memory that associates the CS+ and the US) still exists in the brain after the extinction phase. Indeed, theories of return of fear usually contain the notion that acquisition events and extinction events give rise to separate associations, and thus to separate memories (Bouton, 1993; Gershman, Jones, Norman, Monfils, & Niv, 2013; Craske et al., 2008). These two memories are assumed to compete for retrieval during subsequent presentations of the CS+. At the end of extinction, the extinction memory is stronger, hence the reduction of apparent fear. For spontaneous recovery to occur, the acquisition memory must gain strength over time.

What mechanism drives this change in the fear memory? We propose a novel mechanism of spontaneous recovery: selective maintenance of aversive memories, for example through conscious or subconscious reactivation of these memories. Selective reactivation of memories of negative events may help preserve them in the face of general memory decay and therefore give them a competitive advantage during later retrieval. This idea builds on robust behavioral evidence that emotionally charged experiences are preferentially remembered over time (Dalgleish & Hitchcock, 2023; Rouhani, Niv, Frank, & Schwabe, 2023), as well as neural findings suggesting that memory reactivation can prevent forgetting (Wimmer, Liu, McNamee, & Dolan, 2023).

The main goal of this study is to identify mechanisms that can drive spontaneous recovery of fear. Towards this end, we formulated the selective maintenance idea as a computational model and examined whether the model captures all relevant behavioral features observed in empirical data we collected using a fear conditioning and extinction paradigm. Our experiment consisted of four phases: acquisition, extinction, spontaneous recovery test and relearning, with a short break between acquisition and extinction, and a longer break before the spontaneous recovery test. On each trial, participants saw either a CS+ or a CS-, with the CS+ followed by a loud aversive scream on approximately 50% of trials in the acquisition and relearning phases.

To model the process by which separate memories can be formed for neutral and aversive events associated with the CS+, we used the latent cause framework (Gershman, Blei, & Niv, 2010). This normative Bayesian framework proposes that individuals allocate observations (i.e., combinations of observed stimuli) to different latent causes based on similarity. Latent causes are a formalization of the idea that information can contribute to learning of different associations or contexts – each latent cause is a grouping of past observations that gives rise to predictions about what will happen if that cause is activated again. While the formation of separate memories or associations for aversive and neutral events is necessary for spontaneous recovery, our modeling showed that this was not sufficient and another mechanism was required to cause the aversive or "dangerous" latent cause to gain strength over time. To that end, we added the process of selective maintenance.

Using the latent cause framework, we modeled different variants of selective maintenance, as well as reduced models without selective maintenance and several alternative theories from the literature including recency weighting (Devenport, 1998), neural processing fatigue (Pavlov, 1927) and the enhanced salience of aversive events (see Suppl. Subsec. 2.2 for a discussion of other alternative mechanisms and models). Each model was fit to the behavioral data from each participant in our empirical datasets. We then compared the models quantitatively in terms of how well they predict the empirical data, and in terms of their qualitative ability to generate spontaneous recovery after extinction.

In our exploration dataset, we found that models incorporating selective maintenance provided the best quantitative and qualitative fit to the data. We replicated this pattern of results in preregistered analyses of an unseen dataset from the same study, as well as in data from a separate study with individuals with mental health symptoms. Based on these results, we hypothesized that selective maintenance may be related to initial fear learning and the emergence of anxiety symptoms. We therefore tested for associations between model parameters and anxiety symptoms in both our non-symptomatic and symptomatic samples. Finally, we used the insights gained from our model to propose modifications to exposure therapy that target mechanisms driving spontaneous recovery so as to prevent the return of fear and illustrated the potential effects of such interventions through simulations.

Results

All participants completed an online-administered version of a fear conditioning task (Fig. 1). We first report the results from an exploratory dataset (Study 1, in an undifferentiated population). Subsequently, we report replications of the main results in a second dataset (preregistered analyses on a replication dataset from Study 1) and a dataset from a study with individuals with symptoms of depression (Study 2).

Exploratory dataset (Study 1)

Participants in the exploratory dataset (N=316) learned to differentiate between the CS+ and the CS- during the acquisition phase, and correctly predicted that a scream was likely to follow the CS+ (Fig. 2A, solid black), but not the CS- (Fig. 2A, dashed black). In the extinction phase, participants decreased their expectations of the scream for the CS+, however, there was a marked average increase in expectation of the scream

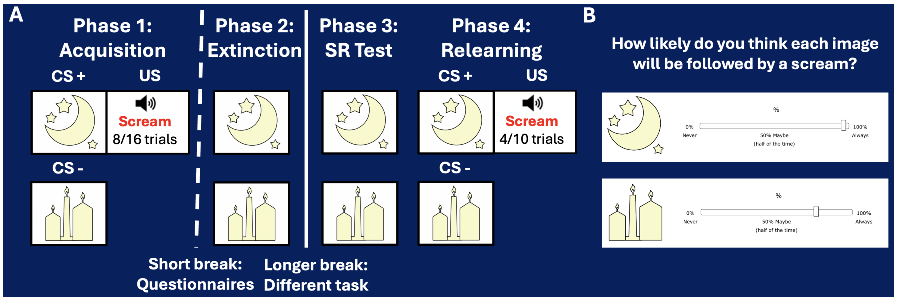
A
Phase 1: Acquisition
CS +
US
Scream
8/16 trials
CS -
Phase 2: Extinction
Phase 3: SR Test
Phase 4: Relearning
CS +
US
Scream
4/10 trials
CS -
Short break: Questionnaires
Longer break: Different task
B
How likely do you think each image will be followed by a scream?
%
0% Never
50% Maybe (half of the time)
100% Always
%
0% Never
50% Maybe (half of the time)
100% Always

Figure 1: **Task design**. **A)** The task consisted of 4 phases. On each trial, participants observed an image of a moon (CS+) or candles (CS-). First, in the acquisition phase, the CS+ was followed by a loud aversive auditory scream (US) on 8 of 16 trials. The CS- was never followed by a US. After a short break (completing a questionnaire for 3-5 minutes), in the extinction phase, no stimulus was followed by the US. After a longer break (completing a different task for 15-20 minutes), a spontaneous recovery (SR) test phase was conducted without any USs. Immediately afterwards, a relearning phase began, in which the CS+ was followed by the US on 4 of 10 trials. **B)** Throughout the task, every few trials, participants were asked to rate how likely they think each stimulus would be followed by a scream (expectancy ratings) on a scale from 0-100%. Occasionally, they were also asked to indicate their emotional response to each stimulus (affective ratings).

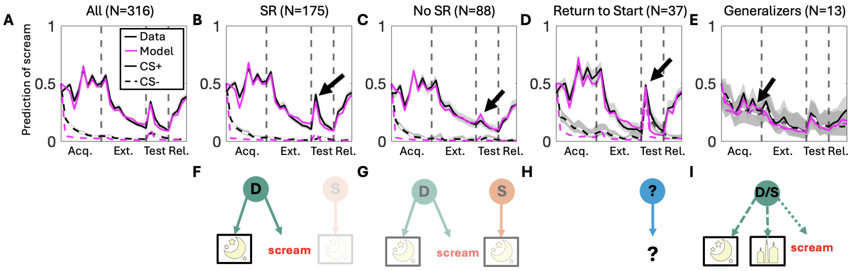

Figure 2: **Empirical expectancy ratings and model predictions throughout the task.** Mean expectancy ratings for the CS+ (solid lines) and CS- (dashed lines) during the four phases of the task (separated by dashed vertical lines; Acq.: acquisition, Ext.: extinction, Test: spontaneous recovery test and Rel.: relearning). Black curves depict the empirical data, pink curves depict simulated predictions from the Selective Maintenance and Low Salience Model averaged across participants. Shading: 95% bootstrapped confidence intervals. **A)** All participants. **B)** Participants who showed spontaneous recovery for CS+ only (spontaneous recovery (SR) group). **C)** Participants who showed no spontaneous recovery (No SR group). **D)** Participants who showed high expectations for both stimuli during the spontaneous recovery test (Return to start group). **E)** Participants who did not differentiate the CS+ from CS- in acquisition (Generalizers group). **F-I)** Illustration of the inferred latent causes that gave rise to the prediction indicated by a black arrow in each group in B-E. 'D' indicates a 'dangerous' latent cause, 'S' a 'safe' latent cause and '?' the creation of a new latent cause.

for the CS+ in the spontaneous recovery test. Finally, participants quickly relearned that the CS+ was likely to be followed by a scream in the relearning phase (Fig. 2A).

Visual inspection of a pilot study (not part of this dataset) suggested that participants' behavior fell into four different subgroups. Simulations with computational models further suggested that the behavior of each subgroup might arise from different mechanisms. To be able to test these mechanistic hypotheses quantitatively, we used the pilot data and simulations to devise a set of criteria to divide participants into four subgroups based on their behavioral response profile (see Data Categorization in Methods). We applied these criteria to our exploratory dataset after the data were collected, to categorize participants into the following subgroups: a spontaneous recovery ('SR') group (participants who showed an increase in CS+ but not CS- expectancy ratings in the spontaneous recovery test; Fig. 2B); a 'No SR' group (participants who did not show an increase in expectation of the scream in the spontaneous recovery test; Fig. 2C); a 'Return to Start' group (participants who showed an increase in expectations of screams for both CS+ and CS- in the spontaneous recovery test; Fig. 2D); and a 'Generalizers' group (participants who did not differentiate between the CS+ and the CS- at the end of acquisition; Fig. 2E). Note that the Generalizers group showed signs of extinction and differentiated CS+ and CS- in the relearning phase, suggesting they were not simply inattentive.

Most participants (all but N=53) fell into the SR (N=175) and No SR (N=88) groups. To examine the mechanisms of spontaneous recovery, we therefore focused our analyses mostly on these two groups. First, we compared the behavior of the SR and No SR groups by analyzing responses for the CS+ (compared to those for the CS-) in the different task phases. Wilcoxon rank-sum tests showed that the SR group had marginally higher expectations of negative outcomes for CS+ in acquisition ($Z = 1.96$, $p = .05$, $r = .12$; Fig. 3A), a larger increase in expectations of negative outcomes for the CS+ after the first break between acquisition and extinction ($Z = 2.03$, $p = .04$, $r = .13$; Fig. 3B), a larger decrease in expectation of negative outcomes for the CS+ during extinction ($Z = 4.93$, $p < .001$, $r = .30$, Fig. 3C); higher differential SR (see below) in the spontaneous recovery test (as expected given that this measure closely related to our categorization criterion; $Z = 13.16$, $p < .001$, $r = .81$; Fig. 3D); and stronger increase of that expectation in the relearning phase ($Z = 4.03$, $p < .001$, $r = .25$; Fig. 3E). Thus, the two groups differed significantly during all phases of the task, and not only in the spontaneous recovery test. Note, however, that the two groups showed no difference in negative outcome expectations before starting the task for each of CS+ alone, CS- alone or the difference between them (all $p > .28$). Medians and interquartile ranges for the comparisons reported here are listed in Suppl. Tab. 1.

Our main aim was to uncover why fear returns with time. To ensure that our behavioral measure of the amount of return of fear was not confounded by processes such as less effective extinction, regression to the mean, or overall decay of memories, we used a measure of "differential SR" throughout our analyses. We computed differential SR as the expectancy rating for the CS+ at the beginning of the spontaneous recovery test (rating at the start of the phase, after the break and before seeing any trials) minus the last expectancy rating for the CS+ in extinction, minus the same metric for the CS-. Differential SR therefore measures the specific change in expectations of a scream for the CS+ during the break, above and beyond any change in expectations for the CS-.

We observed a similar pattern of results when we analyzed the affective ratings using the same approach as for the expectancy ratings (see Suppl. Subsec. 2.1 for details). Given the correspondence between the two measures, and because we had more ratings of expectancies, we focused on the latter for all subsequent computational modeling analyses.

Modeling task behavior to uncover the mechanisms of spontaneous recovery

To understand why spontaneous recovery might arise, we formulated different mechanistic theories (or hypotheses) as computational models and compared them quantitatively (in terms of fit to the empirical data) and qualitatively (in terms of their ability to generate spontaneous recovery).

We note that a central challenge for participants—both in our task and in daily life—is that they do not know the true structure of the environment (e.g., the designation of phases and their boundaries in the task). Instead, they must infer this structure based on past experiences, in order to make accurate predictions. To account for this process of structure inference, we used the latent cause inference framework for all our models. Here, we first describe the building blocks of this framework, and then

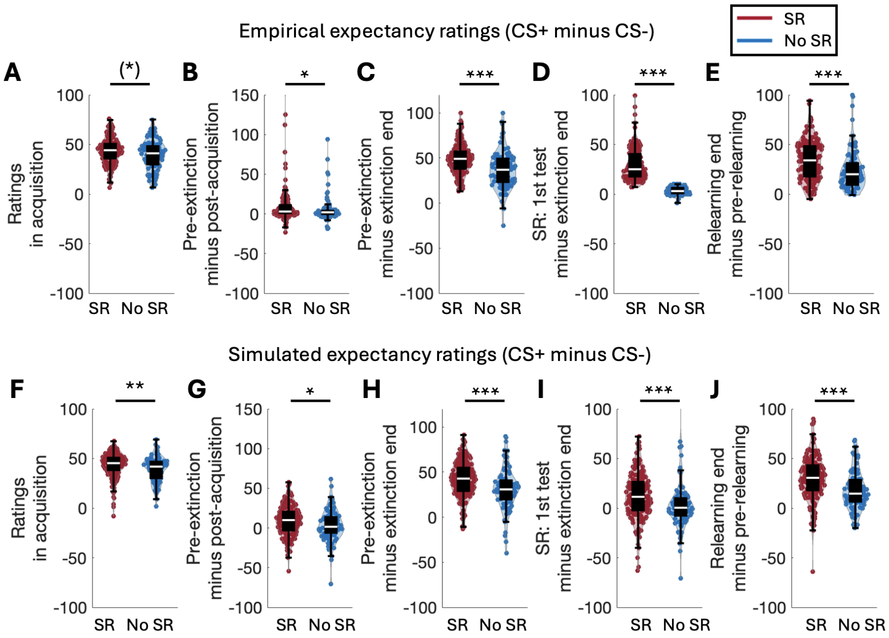

Figure 3: **SR and No SR groups show different behavioral features across all task phases.** Plotted are empirical (**A-E**) and simulated (**F-J**) expectancy ratings for the CS+ minus those for the CS- for different metrics. In red are data of individual participants from the spontaneous recovery (SR) group and blue from the No SR group. White dots indicate the median, the ends of the black boxes indicate the first to the third quartile and the ends of the black lines indicate the minimum and maximum value excluding outliers. **A)** Average ratings during the acquisition stage were higher for the SR group as compared to the No SR group. **B)** Effect of the first break: the difference between the first rating in the extinction phase (prior to any extinction trials) and the last rating in the acquisition phase was higher for the SR group. **C)** Extent of extinction: the difference between the first and last rating in extinction (i.e., the extent of decrease of expectation of the scream during extinction) was larger for the SR group. **D)** Differential spontaneous recovery (SR): The difference between the first rating in the spontaneous recovery test and the last rating in extinction was higher for the SR group (as expected as this comparison closely relates to the criterion for dividing the SR and the No SR groups). **E)** Reacquisition was faster for the SR group, as measured in the difference between the last rating in the relearning phase and the last rating in the spontaneous recovery test phase. Note that the first and last rating in each phase always occurred before any observations and after all observations of that phase, respectively. **F-J)** The same metrics computed from simulated expectancy ratings using the Selective Maintenance and Low Salience Model with the best parameter estimates. *: $p < .05$, **: $p < .01$ ***: $p < .001$.

outline the additional components we included in each model that we fitted to our data (increasing the complexity and adding alternative mechanisms as we proceed). For detailed descriptions and formalizations of each model, see Subsec. 'Generative Models' in the Methods section.

In latent cause inference (Gershman et al., 2010), on each trial the participant infers whether the current observation results from an old (previously inferred) latent cause or a new latent cause. This inference, which is probabilistic (i.e., due to uncertainty about the structure of the task, a trial can be assigned in part to several latent causes), is influenced by at least three factors: 1) the strength of old latent causes, formalized as counts N of how many trials (or fractions of trials) had been inferred to come from each latent cause in the past (the stronger the latent cause, the more likely it will be inferred again), 2) the *a priori* probability of a new latent cause, formalized through a "concentration parameter" α (the larger α the more likely is a new latent cause), and 3) the participant's belief about the variability of observations that each latent cause generates, formalized here (for binary observations) as a beta function with one free parameter a_{obs}. The smaller a_{obs}, the more deterministic (0 or 1) are the probabilities associating the latent cause with each observation, and therefore a new latent cause is more likely to be inferred if a previously observed stimulus is omitted, or vice versa (see Fig. 4A-D for an example).

Inference in this basic latent cause model is not sensitive to time or order of events (Aldous, Ibragimov, Jacod, & Aldous, 1985), and therefore the model cannot account for time-dependent phenomena such as spontaneous recovery. To capture the notion that past events are less relevant for present inferences, we decayed the counts of past events by $\gamma \in [0,1]$ on each trial (Blei & Frazier, 2011). However, simulations show that this decay is not sufficient to explain spontaneous recovery of fear over time, as it does not change the relative strength of different latent causes over time. To address this, we modeled relative *strengthening* over time of latent causes that predict the US, perhaps because they are reactivated in memory. We termed this mechanism "selective maintenance" and formalized it as an increase of the decay parameter γ (leading to less decay) by a selective maintenance parameter ω in proportion to the probability that the latent cause is assumed to lead to an aversive US (Fig. 4E).

Finally, to allow for individual differences in selective maintenance (i.e., some participants may reactivate memories throughout the task while others may do so only during the break), we introduced C_{low}, a low-salience parameter that decreases the counts of events with aversive outcomes when they occur (e.g., during acquisition). During the task, low salience of US events effectively counteracts selective maintenance (which increases the relative strength of latent causes associated with aversive outcomes). However, low salience has no effect during the break, when the effect of selective maintenance can fully unfold. Note that this mechanism was not crucial for explaining spontaneous recovery, but it allowed our model to better capture behavior from a larger variety of individuals. Thus, our final model had four free parameters: prior a_{obs}, decay γ, selective maintenance ω, and low salience C_{low}. We quantitatively compared this model to four reduced versions of itself to test whether each component of the model is necessary.

We also evaluated several models that formalized alternative theoretical accounts of spontaneous recovery from the literature. These included a "Salience Model" that used a salience parameter C for the count for each trial with a US (i.e., in this model a trial with a scream US counts as C trials), to capture the notion that aversive events might be more strongly encoded. We modeled Devenport's (1998) proposal that spontaneous recovery might arise over time due to recency weighting of events in a "Temporal Weighting Model" with a power-law decay of latent-cause counts. Finally, we formalized Pavlov's (1927) suggestion that repeated stimulus presentations reduce processing over time and thus lead to less learning during extinction in a "Processing Loss Model" where the count of each trial was reduced according to the number of times its CS had already been observed. Each of these three alternative models had free parameters governing the processes of interest (see Methods). Overall, this led to eight models, which we also compared to a "Basic α Model" with vanilla latent cause inference and only the concentration parameter as a free parameter. The winning model was reliably recoverable from simulated data when compared to non-nested alternative models (see Suppl. Subsec. 2.4.)

Models with selective maintenance captured the data.

We compared the different models by fitting each model to the trial-by-trial data of each participant and calculating Bayesian Information Criterion (BIC) scores from the likelihood of the data with the best fit parameters, penalized for the number of free parameters. Fig. 5A shows median BIC values and model fits for each model. Signed-rank tests (for non-nested models) and likelihood-ratio tests (for nested models) significantly favored the Selective Maintenance and Low Salience Model over all other models using data from the entire exploratory sample ($p < 0.001$ for all comparisons). Specifically, a signed-rank test showed that the Selective Maintenance and Low Salience Model outperformed the next best non-nested model, the Salience Model (median (IQR): Selective Maintenance and Low Salience Model = 323.18 (282.06–387.89), Salience Model = 327.40 (292.44–395.94); $W = -7.3273$, $p < 0.001$), and a likelihood ratio test showed that it fit the data better than the next best nested model, the Selective Maintenance Model ($\chi^2(316) = 2157.80$, $p < .001$).

Likelihood ratio tests at the level of individual participants revealed that 133 of 316 participants were better fit by the Selective Maintenance and Low Salience Model compared to the Decay and Low Salience Model. Of those, 92 belonged to the SR group and 23 to the No SR group - a significantly disproportionate distribution ($\chi^2(1, N = 263) = 16.63$, $p < .001$, Cramér's $V = .25$), indicating that the model better captured behavior in the SR group.

We had specifically hypothesized the Selective Maintenance and Low Salience Model would explain spontaneous recovery after the long break. To test this, we compared the fit of the 9 models only for participants in the SR group. Paralleling the results from the complete exploratory dataset, the Selective Maintenance and Low Salience Model fit the SR group data better than all other models ($p < .001$, Fig. 5D). Specifically, a signed-rank test showed that the Selective Maintenance and Low Salience Model fit the data better than the next-best non-nested model, the Temporal Weighting Model (median (IQR): Selective Maintenance and Low Salience Model = 315.85 (281.03–382.40), Temporal Weighting Model = 322.99 (292.78–382.16); $W = -5.75$, $p < .001$) and a likelihood ratio test showed that this model fit the data better than

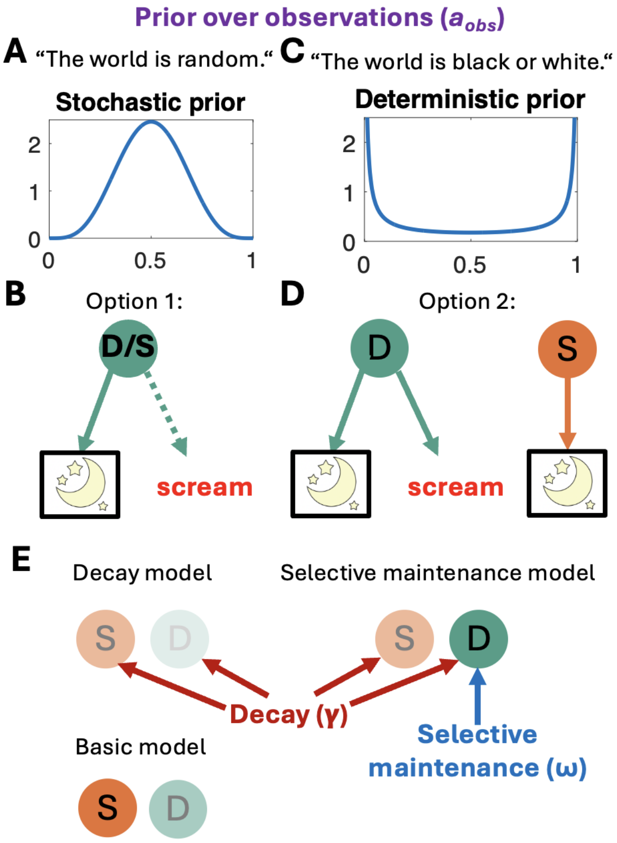

Figure 4: **Model illustration**. Consider a participant who has already assigned the moon (CS+) and the scream (US) to latent cause "D" (for dangerous). If they have a stochastic prior ($a_{obs} >= 1$; **(A)**), when observing a moon without a US they will be more likely to infer this was also generated by D, and update the probability of the US given that latent cause (now D/S – somewhat dangerous and somewhat safe) as their prior allows for probabilistic observations of stimuli given a latent cause **(B)**). In contrast, if they have a deterministic prior ($a_{obs} < 1$; **(C)**), the absence of the US will likely lead to inference of a new safe ("S") latent cause that is only associated with the CS+ **(D)**. The dangerous latent cause will then remain unchanged, despite the safe experience. In **(E)** are plotted relative probabilities of safe and dangerous latent causes at the beginning of the spontaneous recovery test phase, after observation of the moon (a CS+ trial). Darker shade represents a latent cause that is more likely to be inferred. Decay (γ) leads to decreased probability of both safe and dangerous latent causes over time in both the Decay and Selective Maintenance Model. However, in the Selective Maintenance Model (right), the dangerous latent cause is stronger than the safe latent cause due to selective maintenance (ω) that counteracts the decay of (the memory of) that latent cause. In the basic model (left bottom), the safe latent cause is more probable as it was experienced more often.

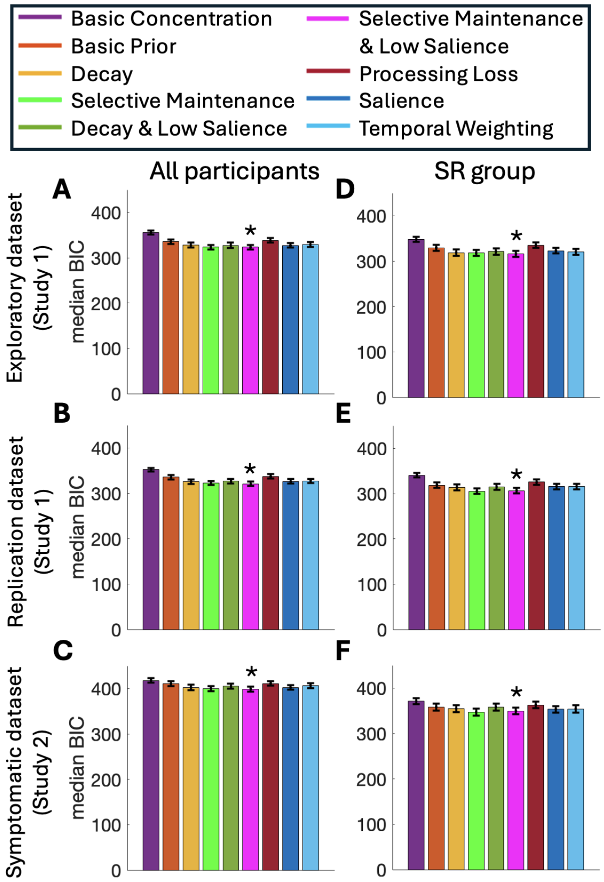

Figure 5: **Models with selective maintenance captured the data best**. **A-C)** median BIC scores for all 9 models fit to all participants in the exploratory dataset in study 1 **(A)**, the replication dataset in study 1 **(B)** and the symptomatic dataset in study 2 **(C)**. Lower BIC scores indicate better model fits. The Selective Maintenance and Low Salience Model (pink) provided the best fit in all three datasets. **D-F)** Median BIC of each model for participants from the spontaneous recovery (SR) group in each dataset. In all three datasets, the Selective Maintenance and Low Salience Model (pink) provided the best fit for this subset of participants as well. Error bars indicate the standard error of the median. * winning model.

the next-best nested model, the Selective Maintenance Model ($\chi^2(175) = 803.05$, $p <$ .001). In contrast, in the No SR group, the Selective Maintenance and Low Salience Model did not explain the data better than the Salience Model (median (IQR): Selective Maintenance and Low Salience Model = 311.45 (274.02–349.79), Salience Model = 310.56 (280.74–350.31); $W = -0.69$, $p = .49$; data not shown).

We focused the remaining analyses of the exploratory dataset on the Selective Maintenance and Low Salience Model given that it captured the data best (the Selective Maintenance Model showed a similar pattern of results).

Parameter comparison between the SR and No SR groups

We formalized three independent hypotheses based on simulations of the Selective Maintenance Model. First, participants who show spontaneous recovery (SR group) should have more selective maintenance of aversive events (as this is what leads to the relative strengthening of the dangerous latent cause during breaks, and therefore to spontaneous recovery of the prediction of the scream). Second, participants whose predictions of screams return to start-of-experiment levels after the long break (Return to Start group) should show a higher rate of decay of events in memory (as this promotes the creation of new latent causes after the break, with their initial prediction levels). Third, participants who fail to distinguish between the CS+ and CS- stimuli during acquisition (Generalizers group) should have more stochastic priors (thus they tend to infer a single latent cause that accounts for observations of both types of stimuli). Due to smaller sample sizes in the Return to Start and Generalizers group, we were not necessarily powered to test the second and third hypotheses and treated these analyses as exploratory.

All parameters of the Selective Maintenance and Low Salience Model were reliably recoverable ($r = .75 - .99$, see Suppl. Subsec. 2.5), and Wilcoxon rank-sum tests supported all three hypotheses. First, estimates of the selective maintenance parameter ω were higher in the SR group than in the No SR group (median (IQR): SR group = 0.49 (0.23–0.73), No SR group = 0.28 (0.01–0.72); $Z = 3.19$, $p = .001$, $r = .20$). This was even more evident when reparameterizing $\omega' = \frac{(1-\gamma)\cdot\omega}{(1-\gamma)\cdot\omega+\gamma}$, which quantifies the proportion of memory of counts due to selective maintenance, which we call 'effective selective maintenance' (median (IQR): SR group = 0.04 (0.02–0.06), No SR group = 0.01 (0.00–0.03); $Z = 5.32$, $p < .001$, $r = .33$, Fig. 6A). Thus, we used effective selective maintenance ω' in later analyses. Second, the Return to Start group had significantly lower decay parameter γ estimates (indicating faster decay; median (IQR): 0.78 (0.66–0.80)) than the SR (median (IQR): 0.91 (0.88–0.94)) and No SR groups (median (IQR): 0.94 (0.90–0.97); $Z = -7.04$, $p < .001$, $r = .48$ and $Z = -6.87$, $p < .001$, $r = .61$ respectively; Fig. 6D). Finally, as shown in Fig. 6G, the prior parameter estimates confirmed that the Generalizers group had more stochastic priors (higher a_{obs}; median (IQR): 0.6 (0–5.99)) than each of the SR group (median (IQR): 1.18×10^{-7} (3×10^{-9}–2.1×10^{-6}); $Z = 2.83$, $p = .005$, $r = .21$) and the No SR group (median (IQR): 5.21×10^{-7} (0–3.5×10^{-5}); $Z = 2.74$, $p = .006$, $r = .27$), and marginally significantly more than the Return to Start group (median (IQR): 5.16×10^{-5} (0–0.11); $Z = 1.84, p = .07$, $r = .26$). Note, however, that comparisons with the Generalizers group must be interpreted with great caution due to the small size of that group

($N = 13$).

We found further support for our hypotheses when inspecting the patterns of inferred latent causes for participants in each group. For participants in the SR group, for the most part, a 'dangerous' latent cause strongly associated with the scream US gained relative strength in the break compared to other, 'safe' latent causes (e.g., those predicting the CS+ but not the US). This accounted for spontaneous recovery of US expectations in the test phase (illustrated in Fig. 2F). Conversely, for participants in the No SR group, both 'safe' and 'dangerous' latent causes decayed over time to same degree (Fig. 2G; see below for a comparison of the strength of the 'dangerous' latent cause between the SR and No SR groups). Furthermore, latent causes inferred for participants in the Return to start group decayed sufficiently over the long break such as to lead to inference of new latent causes in the spontaneous recovery test (Fig. 2H). Finally, the model suggested that participants in the Generalizers group mostly inferred only one latent cause in the acquisition phase, which generated all observed features throughout the task (Fig. 2I).

Confirmatory analyses

Selective maintenance models captured all behavioral features. As can be seen in Fig. 7A, only models that included selective maintenance were able to capture all features of the behavior, in particular, the amount of spontaneous recovery observed in the data (note that all models were fit to the whole sequence of data, without special weighting of the first spontaneous recovery test estimate). Indeed, the Selective Maintenance and Low Salience Model captured all behavioral features of each of the different subgroups (Fig. 2A-E; pink curves), as well as the difference in these behavioral features between the SR and No SR group (Fig. 3F-J).

To verify that the model accounted for all the differences in the behavior of the groups, we analyzed data simulated from the Selective Maintenance and Low Salience Model using the parameters fit to each of the participants in two groups. When comparing data simulated for the SR and No SR groups, Wilcoxon rank-sum tests showed higher predictions of negative outcomes for the CS+ in acquisition ($Z = 3.05$, $p = .002$, $r = 0.19$; Fig. 3F), a larger increase in predictions of negative outcome for the CS+ after the first break between acquisition and extinction ($Z = 2.46$, $p = .01$, $r = .16$; Fig. 3G), a larger decrease of predictions of the negative outcome for the CS+ during extinction ($Z = 4.08$, $p < .001$, $r = .25$; Fig. 3H), more differential spontaneous recovery ($Z = 3.78$, $p < .001$, $r = .23$; Fig. 3I), and a stronger increase of predictions of the negative outcome for the CS+ in the relearning phase in the SR group as compared to the No SR group ($Z = 4.86$, $p < .001$, $r = .30$; Fig. 3J). There were no differences in the simulated predictions of outcomes for each of the CSs in the first trial of the task, before observing any outcomes ($p > .16$ for all comparisons). Thus, the pattern of results from the simulated data matched the pattern of results from the empirical data. Medians and interquartile ranges for these comparisons are displayed in Suppl. Tab. 1.

Importantly, no model without selective maintenance captured the significant differences in differential spontaneous recovery between the two groups ($p > .24$ for all comparisons). Moreover, across participants in both SR and No SR groups, the amount of empirically measured differential spontaneous recovery correlated positively with

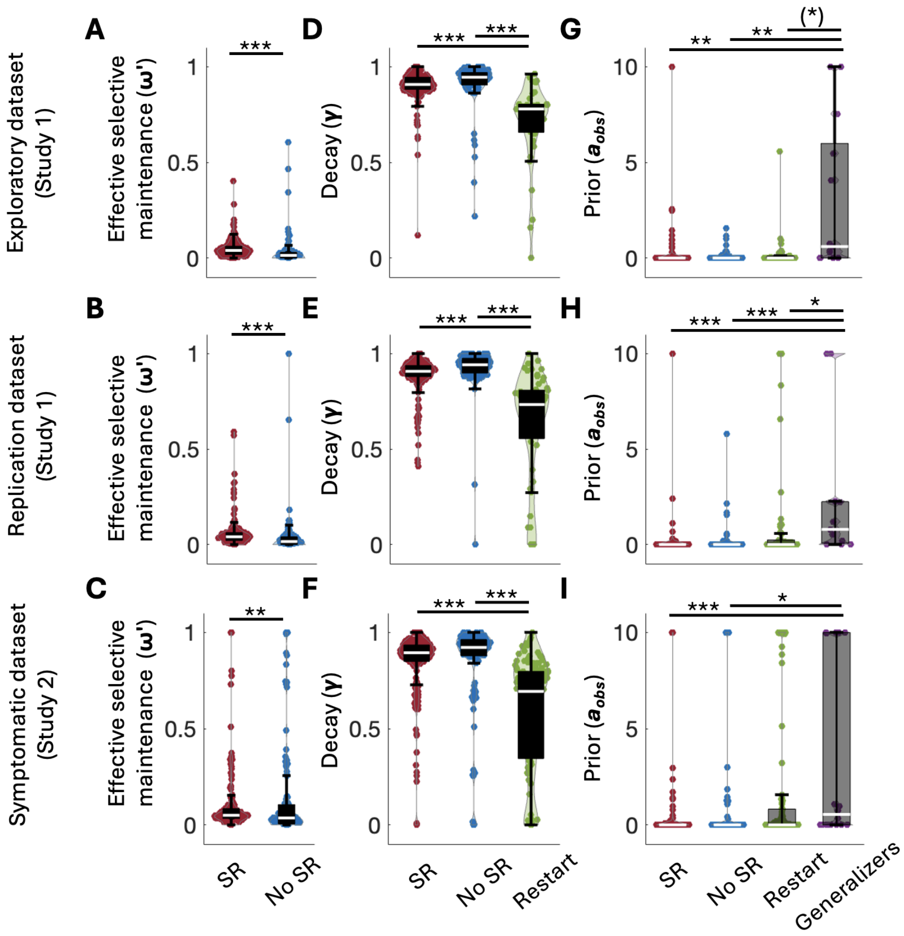

Figure 6: **Parameter comparisons**. Parameter estimates from the exploratory dataset in study 1 are shown in the top row (A, D, G), from the replication dataset in study 1 in the middle row (B,E,H) and from study 2 in the bottom row (C,F,I).**A-C)** Median estimates of effective selective maintenance ω'. **D-F)** Median parameter estimates of decay γ. **G-I)** Median parameter estimates of the prior over observations a_{obs}. Estimates from the spontaneous recovery (SR) group are displayed in red, from the No SR group in blue, from Return to Start group in green and from the Generalizers group in purple. White horizontal lines indicate the group median, colored dots indicate estimates from individual participants, the ends of the black (dark shaded) box indicate the first to the third quartile and the ends of the black line indicate the minimum and maximum value excluding outliers. *: $p < .05$, **: $p < .01$ ***: $p < .001$.

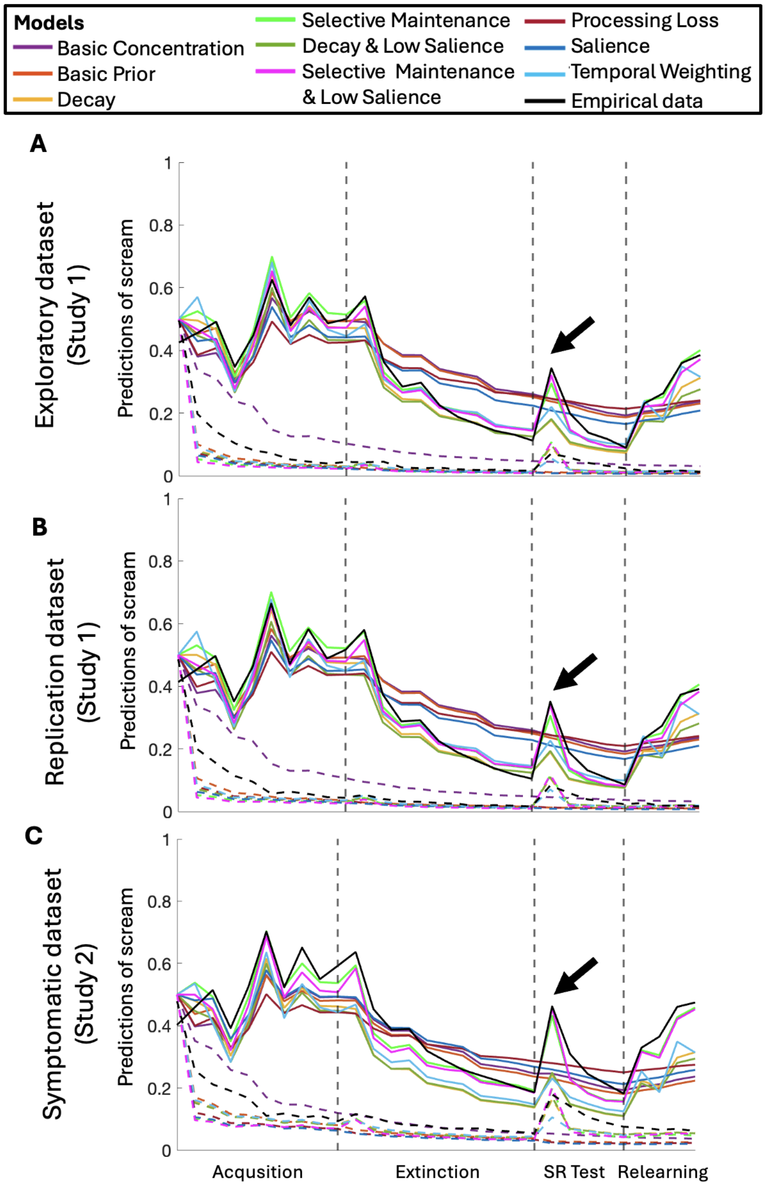

Figure 7: **Models with selective maintenance captured all aspects of the empirical data.** Model predictions of the scream for the moon (CS+, solid lines) and the candles (CS-, dashed lines), simulated using each participant's best-fit parameters for the exploratory dataset **(A)**, the replication dataset **(B)**, and the symptomatic dataset **(C)**. Black: empirical data averaged over all participants. Vertical dashed lines indicate the end of a phase in the task. Only the two models with selective maintenance (light green and pink) quantitatively captured spontaneous recovery at the beginning of the test phase (black arrows; see also Figure 2 for a clearer depiction of the pink model).

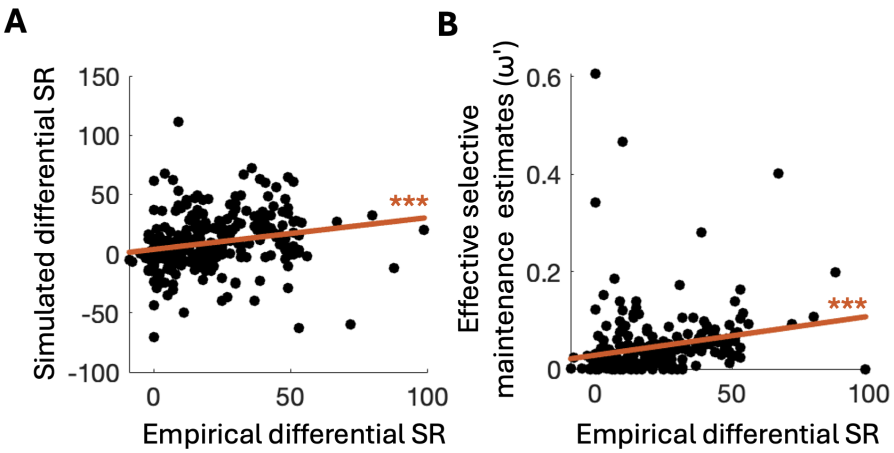

Figure 8: **Selective maintenance as a mechanism of spontaneous recovery.** The amount of differential spontaneous recovery (SR) in the empirical data correlated with the amount of differential SR in data simulated with the Selective Maintenance and Low Salience Model **(A)** and the estimated effective amount of selective maintenance (ω') **(B)** for participants in the SR and No SR groups. *** $p < .001$

differential spontaneous recovery in data simulated using the Selective Maintenance and Low Salience Model ($r = .20$, $p < .001$, Fig. 8A) and with estimates of effective selective maintenance ω' ($r = .22$, $p < .001$, Fig. 8B). Furthermore, the empirical differential spontaneous recovery did not correlate with differential spontaneous recovery in data simulated using models that did not include selective maintenance ($p > .14$ for all correlations).

Spontaneous recovery was driven by the return of the dangerous latent cause. We hypothesized that spontaneous recovery occurs because the strength of the dangerous latent cause becomes higher than that of the safe latent cause associated with the CS+ during the break before the spontaneous recovery test. To examine this, we computed the change in relative strength of each of the two latent causes between the end of extinction and the beginning of the spontaneous recovery test and computed the difference in that change between the two latent causes (dangerous latent cause minus safe CS+ latent cause). Supporting our hypothesis, the difference in change in strength between the two latent causes was larger in the SR group (median (IQR): 0.30 (0.04–0.64)) than the No SR group (median (IQR): 0.01 (-0.01–0.21); $Z = 6.40$, $p < .001$, $r = .40$; Fig. 9A). In line with the assumption that this effect is driven by selective maintenance, effective selective maintenance estimates (ω') correlated with the difference in change in strength between the two latent causes across participants in the SR and No SR groups ($r = .50$, $p < .001$; Fig. 9B). Furthermore, in line with the assumption that these latent cause changes drive spontaneous recovery, this metric of latent

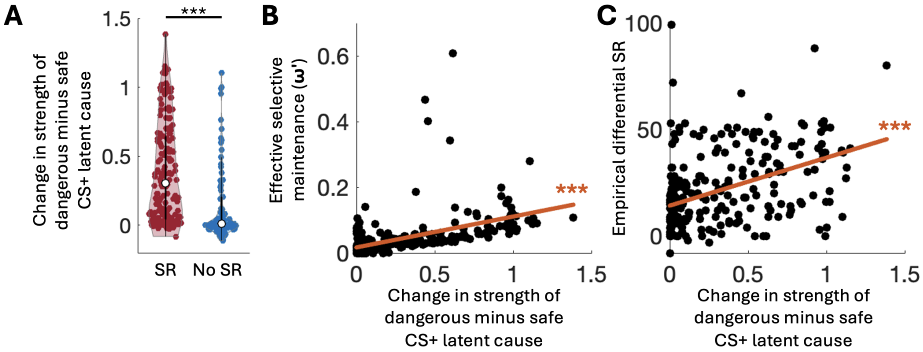

Figure 9: **The strength of the "dangerous" latent cause overpowers other latent causes, resulting in spontaneous recovery.** The relative strength of each latent cause was calculated as the probability of that latent cause being active when observing the CS+ stimulus (before any outcome observation). We computed the change in strength between the end of extinction and the beginning of the spontaneous recovery test for the latent cause that was associated with the CS+ and the US ("dangerous" latent cause) and for the latent cause that was associated only with the CS+ ("safe CS+ latent cause"). **A)** The change in strength of the "dangerous" latent cause minus the change in strength of the "safe" latent cause was larger in the SR group (red) than the No SR group (blue). **B)** The differential change in strength of the latent causes across both SR and No SR groups correlated with effective selective maintenance estimates ω'. **(C)** It also correlated with the empirical differential spontaneous recovery (i.e. change in CS+ ratings minus change in CS- ratings over the same period of time), across both SR and No SR groups. *** $p < .001$.

cause change also correlated with empirical differential spontaneous recovery ($r = .44$, $p < .001$; Fig. 9C).

Results were not driven by differences in response styles. To verify that the differences in ratings between the SR and No SR groups were not driven by differences in participants' response style or usage of the expectancy rating scale, we compared the means and variances of participants' ratings of several probability questions and vignettes. Wilcoxon rank-sum tests did not indicate differences between the groups on any of these measures ($p > .35$ for all comparisons).

Preregistered replication – Study 1

To ensure that our results were robust and we did not overfit our models to our specific dataset, we preregistered a replication of the main results in an unseen replication dataset (N=355) also collected as part of study 1 (preregistration available at `https://doi.org/10.17605/OSF.IO/UCZNF`). Specifically, we preregistered the hypothesis that selective maintenance of aversive memories best explains spontaneous recovery of fear after extinction.

We found a similar distribution of participants into subgroups in the replication dataset with N=172 participants in the SR group, N=110 in the No SR group, N=56 in the Return to start group and N=11 in the Generalizers group. To test our main hypothesis, we conducted two preregistered statistical tests. First, we fit all nine models to the replication dataset. In the SR group, we found that the Selective Maintenance and Low Salience Model explained the data better than the Selective Maintenance model ($\chi^2(172) = 493.37$, $p < .001$), and better than all other models (Fig. 5E). Specifically, the Selective Maintenance and Low Salience Model explained the SR group data better than the next best nested model (i.e., the Decay Model; $\chi^2(344) = 2023.5$, $p < .001$) and the next best non-nested model (i.e., the Salience Model; median (IQR): Selective Maintenance and Low Salience Model = 306.51 (279.59–346.87), Salience Model = 315.58 (293.26–352.80), $W = -7.24$, $p < .001$, $r = .46$). The same pattern of results was seen for the whole replication dataset (Fig. 5B) and the model also captured the whole dataset well (Fig. 7B). Second, we found that the effective amount of selective maintenance was larger in the SR than the No SR group (median (IQR): SR group = 0.04 (0.02 – 0.06), No SR group = 0.02 (0.00 – 0.04); $Z = 5.99$, $p < .001$, $r = .36$; Fig. 6B). Thus, both tests confirmed our preregistered hypothesis.

Additional exploratory analyses also replicated our previous results, showing that the Return to Start group (median (IQR): 0.73 (0.55–0.81)) showed more decay than the SR (median (IQR): 0.91 (0.88–0.94)) and No SR groups (median (IQR): 0.94 (0.90–0.98); $Z = -7.71$, $p < .001$, $r = .51$, and $Z = -8.54$, $p < .001$, $r = .66$, respectively; Fig. 6E) and that participants in the Generalizers group were best fit with a more stochastic prior (i.e., a greater prior parameter; median (IQR): 0.80 (0.07–2.25)) than participants in all other groups (SR group: median (IQR): 2.34×10^{-7} ($5 \times 10^{-10} - 4.42 \times 10^{-6}$)); $Z = 4.15$, $p < .001$, $r = .31$, No SR group: median (IQR): 3.22×10^{-7} ($0 - 1 \times 10^{-5}$)); $Z = 3.96$, $p < .001$, $r = .3602$, Return to Start group: median (IQR): 2.13×10^{-4} ($0 - 0.2408$); $Z = 2.43$, $p < .02$, $r = .30$; Fig. 6H).

Replication of winning model in a symptomatic sample in Study 2

We also applied the analyses from our preregistered replication to a dataset from a symptomatic sample ($N = 520$, comprised mainly of participants with symptoms of depression, i.e., PHQ9 score > 4). In this sample, $N = 209$ were categorized into the SR group, $N = 123$ into the No SR group, $N = 144$ into the Return to Start group and $N = 18$ into the Generalizers group. Similar to the pattern of results in both the exploration and replication datasets in Study 1, the Selective Maintenance and Low Salience Model explained the SR group data better than the Selective Maintenance Model ($\chi^2(209) = 500.75$, $p < .001$) and better than all other models (Fig. 5F). Specifically, it explained the data better than the next best nested model (i.e. the Decay Model; $\chi^2(418) = 33$, $p < .001$) and the next best non-nested model (i.e. the Salience Model; median (IQR): Selective Maintenance and Low Salience Model = 349.59 (305.58–442.33), Salience Model = 353.40 (316.09–437.62); $W = -5.99$, $p < .001$, $r = .41$). The Selective Maintenance and Low Salienace Model also explained the whole dataset from Study 2 better than the other 8 models (Fig. 5C) and captured the data from all participants well

(Fig. 7C). Finally, the effect of selective maintenance was again larger in the SR group as compared to the No SR group (median (IQR): SR group = 0.05 (0.03–0.09), No SR group = 0.04 (0.00–0.11); $Z = 2.61$, $p = .009$, $r = .14$, Fig. 6C)). Exploratory analyses in Study 2 also replicated our previous results, showing more decay of memories (lower decay parameter) in the Return to Start group (median (IQR): 0.69 (0.34–0.80)) as compared to the SR (median (IQR): 0.89 (0.85–0.93)) and No SR groups (median (IQR): 0.92 (0.88–0.96); $Z = -10.44$, $p < .001$, $r = .56$, and $Z = -9.22$, $p < .001$, $r = .56$, respectively, Fig. 6F), and confirming that participants in the Generalizers group were fit as having a more stochastic prior (i.e., larger prior parameter; median (IQR): 0.54 (0–9.99)) than those in the SR (median (IQR): 1.65×10^{-7} (0.0003×10^{-5}–0.36×10^{-5})) and No SR groups (median (IQR): 4.58×10^{-7} ($0 - 1.183 \times 10^{-4}$); $Z = 3.35$, $p < .001$, $r = 0.22$, and $Z = 2.40$, $p < .001$, $r = .20$, respectively; Fig. 6I)).

Associations between symptoms of anxiety and model parameters

Anxiety disorders are highly prevalent. While effective treatments such as exposure therapy exist, symptoms often return even after successful therapy, and anxiety relapses (Craske, Hermans, & Vansteenwegen, 2006). Our model can, at least in principle, explain why and when different anxiety disorders will develop and return. If the mechanisms quantified by our model are causally involved in the development, maintenance, and return of anxiety symptoms, and assuming we can measure these mechanisms with sufficient precision through our task and model parameters, these parameters should be able to predict the course of anxiety in individuals. A randomized controlled study design is needed to fully test such hypotheses. Here, as a first pass, we examined whether these mechanisms are associated with self-reported symptoms of anxiety. Our exploratory analysis was guided by two hypotheses: first, stochastic priors over observations (high a_{obs}) can lead to a spread of fear to many stimuli, and therefore this parameter may be associated with symptoms of generalized anxiety disorder (GAD). Since we had self-reported GAD symptoms in all our datasets, for these analyses we included all participants from each dataset. Second, selective maintenance of aversive events might strengthen the memory or imagination of aversive events and thus drive the development and maintenance of specific anxiety disorders (such as phobias). As selective maintenance could only be reliably estimated in the SR and No SR groups, we only included these subgroups when examining the association between the effective amount of selective maintenance and symptoms of specific anxiety disorders.

Study 1: Symptoms were not associated with model parameters in the general population

The only significant correlation between model parameters and anxiety symptoms in the general population (we tested associations of the prior over observations a_{obs} with GAD symptoms and the effective amount of selective maintenance ω' with symptoms of social anxiety disorder, animal phobia, agoraphobia, situational phobia and traumatic reactions) was a correlation between the effective amount of selective maintenance and traumatic reactions (which measure symptoms of post-traumatic stress disorder; $r = .10$, $p = .03$, $N = 499$). However, this correlation did not survive False

Discovery Rate (FDR) correction for multiple comparisons.

Study 2: Symptoms of generalized anxiety disorder were associated with stochastic priors in a symptomatic sample

Estimates of the prior over observations (a_{obs}) correlated significantly with symptoms of generalized anxiety (as measured by GAD-7 scores) in the symptomatic sample ($r = .15$, $p < .001$). This effect survived FDR correction for multiple comparisons. The prior parameter was also significantly associated with GAD-7 scores in a linear regression model ($F(518) = 11.6$, $p < .001$, adjusted $R^2 = 0.02$, prior: $b = 0.22$, $SE = 0.07$, $p < .001$). This pattern of results remained when controlling for all other parameter estimates and the empirically estimated response noise (σ) in a linear regression model, and when controlling for age, self-reported sex and cognitive ability in the linear regression. In the latter model, being younger and being male were also associated with higher GAD-7 scores ($F(511) = 13.1$, $p < .001$, adjusted $R^2 = 0.09$; age: $b = -0.09$, $SE = 0.02$, $p < .001$; sex: $b = -1.49$, $SE = 0.51$, $p = .004$).

Discussion

Anxiety disorders are widespread, with the most effective treatment being exposure therapy (Parker et al., 2018). Unfortunately, even after successful treatment, symptoms often return over time (Craske & Mystkowski, 2006). In the lab, this phenomenon is seen in spontaneous recovery of fear after extinction training (Rescorla, 2004). Here, we developed and tested a suite of computational models to understand why fear returns with time, and to quantify individual differences that can differentiate those who go on to show spontaneous recovery of fear from those who do not. By fitting different computational models to a large dataset, we showed that spontaneous recovery requires a process that strengthens aversive memories over time – leading us to propose selective maintenance of aversive (from here “dangerous”) latent causes as a mechanism for spontaneous recovery. Models that included such selective maintenance provided the best quantitative fit to three datasets, and were the only models that qualitatively captured the differential spontaneous recovery seen in our behavioral fear conditioning task. Other proposed mechanisms, such as temporal weighting (Devenport, 1998), increased salience of aversive events, or loss of processing of stimuli with repeated exposure (Pavlov, 1927), could not capture differential spontaneous recovery in our dataset. There were significant individual differences in our empirical data, with some participants showing spontaneous recovery of fear, and some showing other patterns of responses. In additional analyses, we showed that the Selective Maintenance and Low Salience Model—the model that best explained participants who showed spontaneous recovery—could also account for the behavior of the other subgroups of participants, including participants who did not show spontaneous recovery of expectations of the scream US over time, participants who generalized across CS+ and CS- stimuli during acquisition, and participants who over time went back to predicting a scream US for both CS+ and CS- as they did at the start of the task. We replicated these results across our exploratory dataset, our preregistered replication in a dataset from the general population, and a dataset from a separate study focusing on participants with symptoms of depression. Supporting our suggestion that the mechanisms identified by our model are relevant for anxiety disorder, we showed that a model parameter quantifying the degree to which participants tend to generalize across the CS+ and CS- in our task correlated with symptoms of generalized anxiety disorder in the latter dataset. As we will illustrate below, the model makes a set of predictions on how we can improve exposure interventions to prevent the return of fear.

Generalization of our results to affective ratings

A major goal of our research is to understand how we can improve psychological interventions such that an individual feels better according to their own subjective experience and report. In our experiment, we elicited both explicit expectation of the scream US (which were analyzed in the results presented here), as well as affective ratings for each of the CSs. Importantly, we saw a similar pattern of responses for affective ratings as we did for expectancy across all four groups. This included not only spontaneous recovery after extinction, but also significant differences in affective ratings between the SR and No SR groups throughout the phases of the task (Suppl. Fig. 2), similar to what we found for expectancies (Fig. 3). As such, we believe our findings are not limited to explicitly elicited expectancies, but also to emotional processes such as fear. However, the degree to which this pattern of results also generalizes to physiological measures remains to be investigated.

Improving interventions using model-driven mechanistic insight

Considering that extinction and spontaneous recovery are models of exposure therapy and relapse in anxiety disorders, we propose that the mechanistic insight gained from modeling such behavior can be used to improve interventions of exposure therapy. Specifically, the model suggests that intervening in any of its building blocks—deterministic priors, memory decay, or selective maintenance of aversive memories—could help prevent spontaneous recovery.

Two interventions proposed in the literature have already been empirically shown to reduce the amount of return of fear. These are: spreading extinction trials (exposure sessions) over longer time intervals (Laborda, McConnell, & Miller, 2011) or tagging the ‘safe’ (exposure) memory with a retrieval cue such as a wristband that clients can then wear and use as cue to recall that memory (Craske, Treanor, Conway, Zbozinek, & Vervliet, 2014). Our model can explain why these interventions work. Specifically, it suggests that these interventions strengthen the safe memory and thus counteract selective maintenance of ‘dangerous’ memories. While the model predicts these interventions prevent dangerous memories from spontaneously resurfacing over time, the ‘dangerous’ memory may still resurface if it is triggered directly.

The model’s predictions also give rise to two novel interventions to eliminate the effect of selective maintenance altogether (and thus prevent relapse). First, cognitive restructuring focusing on reducing “black-and-white thinking” prior to exposure therapy may change deterministic beliefs (i.e., increase α_{obs}) and help prevent inference of a new latent cause during exposure therapy (Smith et al., 2021). Second, selective maintenance of aversive memories might result from negative self schemas (Dalgleish & Hitchcock, 2023), which may bias sampling from memories towards aversive events. Schema therapy (Young, Klosko, & Weishaar, 2006) offers a host of interventions to reduce negative self-schemas, including cognitive, emotional and behavioral techniques. Addressing negative self-schemas prior to (or even after) exposure might also prevent later return of fear.

To test our prediction that interfering with any of the mechanisms involved in spontaneous recovery (possibly through the above mentioned interventions) can causally reduce spontaneous recovery, we manipulated the influence of each mechanism in turn by changing parameter values in line with how we hypothesize they would be changed by the proposed interventions, and simulated new data to examine whether spontaneous recovery would in fact be reduced. Fig. 10 shows such simulations for one example participant from our exploratory dataset, demonstrating that spontaneous recovery could be prevented by all four interventions.

We would like to note that we outline here a set of hypotheses on how interventions may be improved, which can serve to inform new empirical studies. However, empirical support of several of these ideas is still lacking and we do not recommend their implementation during treatment prior to rigorous testing (but see Craske, Treanor, Zbozinek, and Vervliet (2022) for an overview on how to improve exposure therapy).

Extension to other return-of-fear phenomena

Two other return-of-fear phenomena are often observed after fear conditioning and extinction. These include the return of fear when returning to the acquisition context (if acquisition and extinction occurred in different contexts) or when switching to a completely new context ('renewal'), and after re-exposure to the US alone ('reinstatement'). Our selective maintenance model expands previous theories put forward to account for these phenomena.

Renewal refers to the return of fear after context switches due to (observable) cues (e.g., background color or location; for temporal contexts see the Suppl. Subsubsec. 2.2.2). Usually, one context is used in acquisition (context A), one in extinction (context B) and finally participants' fear levels are tested in context A, B, or C. Context theories (Bouton, 1993, 2004) explain the general pattern of higher fear in context A, no return of fear in context B and a weighted average amount of fear in context C. However, fear responses in context C that exceed the weighted average of fear shown in contexts A and B, for which there is evidence in animals, but not in humans (see meta-analyses by Wang et al., 2024; Effting et al., 2007), are difficult to explain using context theory alone. To explain this, context theories suggest that only extinction is context-dependent as it occurs later in training, and therefore a new context generalizes more strongly to the non-context-specific fear acquisition phase. However, in daily life situations, it is most likely that people were first exposed to the stimuli they later learned to fear without negative outcomes (otherwise, if the feared stimuli were so rarely encountered, they would not cause debilitating anxiety and exposure therapy would not be necessary). As such, fear acquisition would also be a later learning phenomenon that would be tied to a context, therefore generalization to a new context

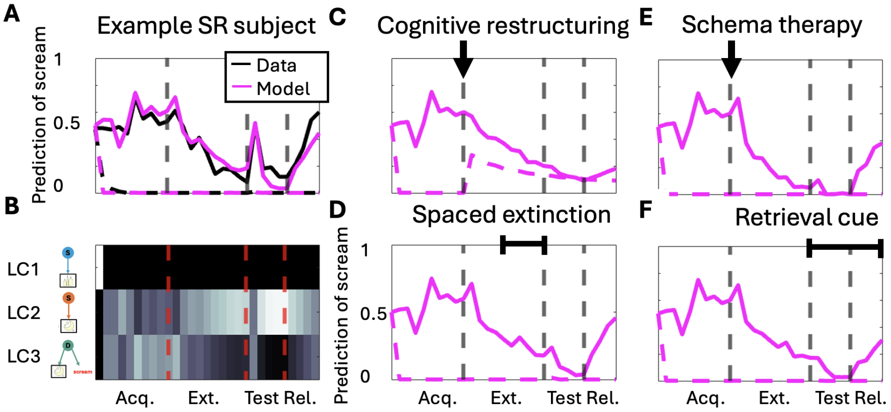

Figure 10: **Simulated intervention effects**. **A**) Empirical data (black) of a single participant who showed spontaneous recovery (SR) and simulation (pink) of the Selective Maintenance and Low Salience Model using the best-fit parameters for this participant. **B**) Latent causes inferred by the model at each prediction of the US after seeing the CS+. This inference gives rise to the prediction in the pink solid line in A. In acquisition, latent cause 1 (LC1) was associated with the CS-, LC2 with the CS+ and LC3 with the CS+ and the US. Lighter gray indicates a higher probability that the latent cause is active on each prediction trial. LC3, whose probablility was reduced through extinction, becomes more likely between the end of extinction and the spontaneous recovery test due to selective maintenance. **C**) We simulated the effect of cognitive restructuring by increasing a_{obs} (setting it to 3.5) to decrease “black-and-white thinking” before exposure (extinction). **D**) We simulated spaced exposure by moving some extinction trials from the second half of extinction into the break phase. **E**) The effect of schema therapy before exposure (extinction), simulated by setting the selective maintenance parameter to 0. **F**) We simulated a retrieval cue after exposure (extinction) by adding selective maintenance to LC2 and LC1 (‘safe’ latent causes) such that decay rates for all latent causes were $\gamma + (1-\gamma) * \omega$, previously applied only to aversive memories. In all plots, dashed vertical lines indicate the last prediction of a phase. Solid lines: CS+ predictions. dashed lines: CS- predictions. Black symbols indicate when the intervention was applied. In all cases, the manipulation decreased spontaneous recovery.

would not necessarily overemphasize this stage. In contrast, selective maintenance of aversive events can explain renewal independent of whether preexposure to the CSs without the USs occurred or not, as selective maintenance would enhance the relative strength of aversive memories from context A because they are aversive.

Reinstatement refers to the return of fear after the occurrence of an unpaired US. One explanation is that reinstatement occurs because the US observation leads to the inference of the CS-US latent cause, in line with predictions of our model. However, the process of selective maintenance suggests an additional pathway. Negative experiences (e.g., with the US) may trigger or enhance selective maintenance, for instance through reactivation of other negative memories or the activation of negative core beliefs that bias sampling from memories. Thus, for some individuals, selective maintenance may only be active (or may be especially active) during periods of stress/negative affect such as are triggered by experiencing negative events. This could help explain why relapse rates tend to be higher during stressful times.

Mechanistic insights from the model explain phenomena of anxiety

Overgeneralization in anxiety. A commonly observed phenomenon in individuals with anxiety disorders is overgeneralization (Cooper et al., 2022), i.e., generalization of learned expectations from a few negative experiences to many (new) situations and contexts. Based on our computational framework, we propose that there are at least two possibly distinct types of overgeneralization. The first type involves a failure to distinguish between stimuli. This resembles the behavior observed in the Generalizers group, resulting from the inference of mainly a single latent cause that accounts for different observations. In a previous experiment in which we tested generalization of US predictions to (previously unseen) stimuli spanning the perceptual range between the CS+ and the CS- (Aitsahalia, 2022), we indeed showed that the behavior of participants from the Generalizers group in the generalization test closely resembles the pattern of 'overgeneralization' proposed by Cooper et al. (2022).

Inferring a single latent cause for distinct stimuli could lead to fear of many situations and stimuli also associated with that same latent cause. (In principle, this type of overgeneralization does not need to be biased to negative experience, but could be in case of increased salience of aversive events.) Interestingly, behavior resulting from the inference of a single latent cause (due to stochastic priors in the Generalizers group) aligns with the main criterion of generalized anxiety disorder according to the DSM-5: "Excessive anxiety and worry (apprehensive expectation), occurring more days than not for at least 6 months, **about a number of events or activities**" (emphasis ours). In line with the model's prediction that stochastic priors can drive generalized anxiety symptoms, in our symptomatic sample we found that participants who were fit with more stochastic priors tended to have more symptoms of generalized anxiety disorder.

We note that for people showing this type of generalization, extinction is expected to occur slowly but should eventually be successful. Although this extinction is likely to be resistant to the return of fear, its gradual pace may be misinterpreted as a lack of response to exposure therapy, thereby resulting in its premature termination. Our model predicts that exposure therapy will show its effects faster when latent causes are more distinctly separated. Thus, to facilitate separation, interventions can promote attention to differences between stimuli or situations, such as is done in interpersonal discrimination exercises (McCullough, Schramm, & Penberthy, 2014).

A second alternative type of overgeneralization that is specific to negative events can simply result from selective maintenance of such events. Selective maintenance would make the latent causes (memories) that involve aversive events stronger with time, such that they carry more weight when an individual is confronted with an even remotely similar situation. That is, due to the strength of the latent cause, it will be *a* *priori* likely to be inferred in many different situations.

The development of anxiety. In line with suggestions in the literature (e.g., Gagne, Dayan, & Bishop, 2018), selective maintenance of aversive experiences in memory can explain the development of anxiety disorders (including post traumatic stress disorder) after an exceptionally aversive event due to increased reactivation of the memory of the aversive ("dangerous") event. With time, the aversive memory overshadows "safe" memories of similar situations without aversive outcomes, even if the safe events had occurred more frequently than the dangerous event. Avoidance behavior further facilitates the memory imbalance in favor of the "dangerous" memory, as it prevents new non-aversive experiences that would strengthen the "safe" memory. As such, a vicious cycle of reactivation and avoidance can lead to the development of an anxiety disorder or post-traumatic stress disorder. This prediction aligns with our exploratory finding that symptoms of traumatic reactions were correlated with larger estimates of the effect of selective maintenance in our undifferentiated sample (note, however, that this result did not survive correction for multiple comparisons and thus needs to be interpreted with caution and further tested, ideally with a more symptomatic sample).

Intolerance of uncertainty in anxiety. Difficulties in tolerating uncertainty have recently been proposed as a theory of anxiety. Brown, Price, and Dombrovski (2023) reviewed a series of studies suggesting that deficits in dealing with uncertainty are more prevalent in anxiety than are exaggerated fear of aversive events. This aligns with the effect of selective maintenance of aversive memories, which can provide a causal explanation for how aberrant behavior in the face of uncertainty comes about in individuals with anxiety. The idea is that uncertainty about the occurrence of aversive outcomes will have led to forming both a "safe" and a "dangerous" latent cause. Selective maintenance then increases the relative strength of the "dangerous" latent cause, leading to higher estimates of aversive outcomes in such situations, similar to the overestimation of the probability of a scream in acquisition that we saw in the SR group. In contrast, in situations with certain aversive outcomes, selective maintenance would have no effect and would not predict different behavior for individuals with anxiety.

Limitations and potential extensions of our study

Alternative models. The potential model space and combination of mechanisms is, of course, endless. For clarity and comprehensibility, we reported here a set of models that represent either key distinct hypotheses from the literature or our own hypotheses that we considered biologically and psychologically plausible. In the Suppl. Subsec. 2.2, we provide an overview of further alternative modeling frameworks (reinforcement learning), theories (temporal context and response fatigue) and other versions of our own models and describe why they fail to account for our data. These alternative models were either unable to generate differential spontaneous recovery or they did not capture the empirical data better than the Selective Maintenance and Low Salience Model. Nevertheless, we cannot exclude that other formalizations or interpretations of these or other alternative mechanisms can account for spontaneous recovery of fear.

For example, our current formalization and data cannot differentiate between selective maintenance of aversive events and faster decay rates of inhibitory associations as proposed by Paskewitz, Stoddard, and Jones (2022). Indeed, it is likely that different mechanisms are at play in different individuals and that the observed behavior results from a combination of mechanisms.

We note that our conclusions are also limited by the (relatively) small amount of data we had in two aspects: First, we had few data points (ratings) per participant and used only one data type (expectancy ratings). Because we used principles of parsimony that seek to identify the simplest model that explains the data (i.e., we penalized models for their free parameters; Wilson & Collins, 2019), we may have missed more complex processes (e.g., additional temporal organizations of context) that would only become visible when additional data, for instance from direct measurements of memory, are also assessed. Second, we had only one trial per participant to measure spontaneous recovery. This not only makes it challenging to estimate spontaneous recovery reliably, but also poses a challenge for model comparison because the single trial can only provide limited support for models that specifically capture spontaneous recovery. In this respect, the fact that models with selective maintenance won when penalized for extra parameters shows that this process was important not only for spontaneous recovery, but rather had signatures throughout the learning process.

Improving measurement accuracy. While our findings were robust and replicated across datasets, most effect sizes were small to medium. Additionally, model fits at the individual level revealed that our model did not fully capture the extent of spontaneous recovery in all participants. This may be due to different mechanisms underlying spontaneous recovery across individuals, as previously discussed. Alternatively, the variability could result from noise in the data or limitations in the fitting procedure, which may not always recover optimal parameter estimates given the stochastic nature of the model. Increasing the number of trials and applying hierarchical modeling approaches could improve the robustness of parameter estimation, thereby enhancing the precision of measuring these mechanisms in individuals.

Alternative task design. While our pre-registered replication and additional replication in a symptomatic sample give confidence that our model was not overfit to a specific dataset, it could nevertheless be too specific to our task design and thus not generalize to other versions of this paradigm. For instance, the amount of spontaneous recovery could be triggered by the specific sequence of reinforced and non-reinforced trials, which was identical across participants. Reassuringly, we previously observed similar patterns of behavior in other versions of the task with different numbers and sequences of trials in acquisition and extinction, as well as with longer intervals during the break (e.g. testing the following day; dataset not shown). One main result from these different variants of the task is that longer breaks increased the number of participants in the Return to Start group, who tend to rate expectations of the scream at 50% for both the CS+ and the CS- in the spontaneous recovery test. This is in line with our model's prediction that over time, memory decay can override selective maintenance, especially in a task such as ours, where aversive memories are stronger than neutral ones but not overwhelmingly dominant. Formally, our model accounted for this phenomenon in the data from the Return to Start group either through the formation of new latent causes for the CS+ and the CS- in the spontaneous recovery test, or through a return of the "dangerous" latent cause for the CS+ and the formation of a new latent cause for the CS-. Data from the present paradigm cannot tease these different latent cause inferences apart, which is why we excluded data from the Return to Start group from analyses on selective maintenance.

Conclusion

We propose selective maintenance of aversive memories as a key mechanism underlying spontaneous recovery of fear after extinction. Models incorporating this process explained the behavioral data across multiple datasets, outperforming alternative theories. These models also provided mechanistic insights into overgeneralization and the development of anxiety. These insights were empirically supported by correlations between parameter estimates quantifying the proposed mechanisms and symptoms of anxiety.

These results may have clinical implications: targeting mechanisms supporting selective maintenance, for example through cognitive interventions or spaced extinction, may help reduce the risk of relapse after successful treatment. While further research is needed to test these clinical predictions and the model's neural basis, selective maintenance of aversive memories offers a unified, testable account of the development, persistence and return of anxiety, and how the influence of selective maintenance might be effectively modified.

Data availability

We will make all raw and modeled task data from participants included in the analyses in this manuscript available in a data repository upon publication of this manuscript. Mental health symptom data will not be made available publicly due to concerns that full anonymization of such data may be impossible given our current stage of knowledge (and future algorithms). More broadly, we are concerned about potential misuse of such data to disadvantage people with mental health symptoms. We will make these data available to researchers upon request within two weeks after contacting the corresponding author if they provide a convincing plan to use the data in a way that does not endanger individuals suffering from mental health symptoms and can ensure that the data will be kept secure. We will make all data available to editors and reviewers during the review process.

Code availability

We will make all code of our computational modeling pipeline and all code related to our main hypotheses available as a git repository upon publication of the manuscript and to editors and reviewers during the review process. Specifically, we already deposited the code for data preparation, computational modeling and statistical analyses of our main and preregistered hypothesis to a git repository.

Acknowledgement

Isabel Berwian, Yongjing Ren, Jialing Ding, Seohyun Moon and Yael Niv's work on "Precision psychiatry for treatment selection in depression" was supported by Wellcome Leap as part of the Multi-Channel Psych Program. Research reported in this publication was in part supported by the National Institute of Mental Health of the National Institutes of Health under award P50MH136296. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We would like to thank Sebastian N. Quinard for creating the auditory components of the task and Augustin Hennings for discussion on its relation to memory research.

Conflicts of Interest

The authors report no conflicts of interest.

Method

Procedure and participants – Study 1

This study was approved by the Institutional Review Board of Princeton University (Protocol 11968), and all participants provided written informed consent. Participants were recruited from Prolific to complete a series of studies including online behavioral tasks and mental health questionnaires in spring 2023. All participants were compensated for their time at a rate of $13/hr. To qualify for the study, participants had to reside in the United States, Canada, Australia, or New Zealand, be fluent in English, and have headphones. We first report the analyses of the first half of that dataset with $N = 376$ participants, which were labeled an 'exploration dataset.' (Later, we will report replication of our findings in the second half of the dataset.) From the exploration dataset, 23 participants were excluded due to incomplete datasets and 37 participants were excluded due to failing more than two audio attention checks. The final dataset contained $N = 316$ participants.

Behavioral Task

We designed an online-administered aversive conditioning and extinction task (Fig. 1). On each trial, participants saw a face-down card that then "flipped over" to reveal one of two stimuli: a moon with three stars (CS+) or three candles (CS-). Participants then had to press the space bar to reveal the trial's outcome, described in the instructions as "the sound associated with that stimulus". On some CS+ trials, this was an aversive but tolerable auditory scream (US; see below), and for all other trials there was no outcome. The task was completely Pavlovian; key pressing allowed reaction-time measurements and ensured continued attention to CSs, but did not otherwise affect the occurrence of the US.

The sequence of a trial was as follows: a card appeared face-down for 500 ms and then flipped over (this animation lasted 100 ms). An event listener was then activated and the participant had 6000 ms to press the space bar before the trial timed out. If the participant failed to respond in time, a warning message was displayed for 3000 ms, after which the trial was repeated. If the space bar was pressed and the stimulus had an associated sound, a delay of between 300-800 ms occurred before the audio was played. This jitter was randomly generated on each trial and was introduced to prevent participants from precisely anticipating the timing of the outcome. After the space bar was pressed, the card flipped back over and flew off the screen to indicate the end of the trial (this animation lasted 500 ms, regardless of whether a sound was played). We used three aversive audio clips across trials. They were edited to be of similar volume and duration (2000 ms).

Every three trials, on average, participants were asked to rate how likely they expected each of the two CSs would be followed by a scream on a scale of 0-100% using two horizontal slider bars. The sliders were initialized in a random location and each had to be moved to continue to the next trial. Participants completed 31 ratings for each CS. Within these, two sets of ratings that followed previous ratings without new trials or significant passage of time since the previous rating were excluded from analyses, leaving 29 ratings per CS per participant. One rating was always completed before each phase. In addition, about every 8-10 trials, participants rated their feeling when viewing each of the stimuli on a horizontal slider bar ranging from calm to alert ("affective ratings" hereafter).

The task consisted of 4 phases. 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. This phase was followed by a 3-5 minute break in which participants completed a filler questionnaire of no interest (the 20-item short form International Personality Item Pool-Five-Factor Model measure; Donnellan, Oswald, Baird, & Lucas, 2006). The extinction phase included 30 trials, 18 CS+, 12 CS-, with no US. Next, participants completed a separate task (a risky decision making task in which they made choices between a sure 5-point reward and a stimulus with an unknown but learnable chance of giving 10 points or 0 points; not analyzed here) for $\sim$15 minutes. This was followed by a spontaneous recovery test (16 trials, 8 each of CS+ and CS-; no US), and then a relearning phase (10 CS+ with 4 USs; 6 CS-). All participants saw the same sequence of CSs and USs and were asked to provide ratings between the same trials (see Suppl. Sub. 1.1 for the order of trials and other technical details). As participants were tested online and from remotely, to ensure they heard the US loudly enough and did not reduce the volume throughout the task, participants completed a set of volume calibration steps and six auditory attention checks were spread throughout the task (see Suppl. Subsec. 1.1). Including the two breaks, the task took approximately 50 min.

Self-report measures

Anxiety questionnaires

In a separate online session, participants completed the GAD-7 questionnaire (Spitzer, Kroenke, Williams, & Löwe, 2006; Löwe et al., 2008), several subscales of the HiTOP questionnaire (Watson et al., 2022), and other symptom self-report questionnaires not analyzed here. We analyze here data from the social anxiety subscale (15 items), the agoraphobia subscale (6 items), animal phobia (6 items), situational phobia (6 items), traumatic reactions subscale (8 items). See Suppl. Subsec. 1.2 for a list of items for each subscale. These items were selected based on the recommendations of the HiTOP consortium based on preliminary testing (Watson et al., 2022) and to maximize content overlap with several standard social anxiety scales (Mattick & Clarke, 1998; Heimberg et al., 1999). Participants were invited to complete the HiTOP questions up to several weeks after completing the task. However, we note that participants were instructed to indicate for each statement "Have there been significant times during the last 12 months in which the following statements applied to you?" which also covered the period of completing the task.

Response style assessments

To verify that behavioral differences in the task were not driven by different usage of the expectancy rating scale (i.e., differences in people's response style), we added two types of response style measures. First, we created five anchoring vignettes, in which we described the number of neutral and aversive stimuli observed by another person and asked participants to indicate how that person should rate the expectancy of negative outcomes (Hopkins & King, 2010). Second, we asked participants to rate the likelihood of four events such as "How likely is it that the sun will rise tomorrow?". The scenarios were chosen such that no expert knowledge was required or would influence the rating, and such that ratings would represent participants' responding across the range of probabilities. One goal was to test whether some participants did not use the whole scale even for expectations of "never" or "always". For both measures, we computed the mean and variance of each participant's ratings. All questions are available in Suppl. Subsec. 1.4.

Data categorization

In our pilot data, we observed that most participants showed distinct patterns of behavioral features that suggested they could be categorized into different groups. In parallel, simulations with our model prior to fitting it to any data also indicated that these behavioral features can arise from different mechanistic functions. To provide statistical evidence for the hypotheses emerging from simulations, we therefore decided on the below criteria to separate individuals into four subgroups *a priori* and included the same criteria in our preregistration of the replication dataset.

Specifically, we categorized participants into four subgroups: $N = 13$ participants who showed similar expectations for the CS+ and CS- at the end of acquisition (< 10 points difference on a scale from 0 to 100%) were categorized into the 'Generalizers' group. The remaining participants were then categorized into groups based on their behavior on the spontaneous recovery test, in particular, their first rating of expectation of a scream for both the CS+ and the CS-, which occurred immediately at the beginning of the spontaneous recovery test phase before any other exposure to the stimuli or screams or audio instructions. $N = 175$ participants who increased their expectation of the US by more than 10 points compared to the end of extinction for the CS+ but not for the CS- were categorized into the 'spontaneous recovery' ('SR') group—a group that includes participants who show *differential* spontaneous recovery, i.e., an increase in their CS+ rating without a substantial increase in their CS- rating in the spontaneous recovery test. $N = 88$ participants whose expectations for both stimuli increased by at most 10 points were categorized into the 'No SR' group (we used a change in expectation of 10 points as a cutoff for distinguishing between No SR and SR to account for random response noise), and $N = 37$ participants whose expectations increased by more than 10 points for both stimuli were categorized into the 'Return to Start' group. $N = 3$ participants who did not meet any of these criteria were not included in subgroup analyses.

Model-agnostic analyses

Most of our analyses focused on the SR and No SR groups, as differences between these groups were of main interest to understand the mechanism of spontaneous recovery, and these were also the largest groups, which allowed us to identify robust group differences. First, in a set of model-agnostic analyses, we tested whether the SR and No SR groups showed different behavioral features throughout the task (beyond the differences in the SR test phase, by which these groups were defined) by comparing 1) their *a priori* expectations of negative outcomes for the CS+ and CS- (i.e., their first expectancy ratings); 2) their acquisition learning, measured as the mean differential ratings for the CS+ minus CS- in the remaining ratings from the acquisition phase; 3) the effect of the first (short) break, by comparing the change in expectancy ratings for the CS+ minus CS- in the first rating after the break minus the last rating before the break; 4) the extent to which expectations of the US decreased during the extinction phase in each group, measured as the difference between the first and last ratings for the CS+ minus the CS- in the extinction phase; 5) their differential spontaneous recovery, by comparing the change in CS+ minus CS- ratings between the end of extinction and the beginning of the spontaneous recovery test; and 6) their rate of reacquisiton, by comparing the change in CS+ minus CS- ratings from the last rating of the spontaneous recovery test phase to the last rating of the relearning phase.

We repeated these analyses for the affective ratings with the exception of the effect of the first short break as we did not measure affective ratings immediately after that break. For a similar reason, we also used the affective rating after acquisition rather than the first rating of extinction to measure the effect of extinction (number 4 above).

Generative models

Next, we formalized learning of the latent structure of the task using Bayesian non-parametric models of latent cause inference (Gershman et al., 2015; Gershman & Blei, 2012). In the latent cause framework, observations (here, trials) are attributed 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' (to model the time between acquisition and extinction and between extinction and spontaneous recovery test) and i=4 to indicate the US. Thus, in our models, 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 use a Chinese Restaurant Process (CRP) prior (Aldous et al., 1985) that expects latent cause L_{t+1} on trial $t+1$ to be a previously encountered or a new latent cause with probabilities as follows:

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

where $N_{j,t}$ is the number of observations generated by latent cause number j up to trial t (see more below) and $\alpha \geq 0$ is a parameter that influences the probability of new latent causes. (Note that there is some non-zero probability of a new latent cause on every trial, hence the number of previous latent causes on trial $t+1$ is t. Also, while in some models $\Sigma_{m=1}^{t} N_{m,t} = t$, this is not the case in all our models below, hence we spell out the sum explicitly in the denominator.)

Prior over observations

Conditional on latent cause j having been drawn for trial $t+1$, the generative model assumes that observations $\mathbf{O}_{t+1}$ are generated by independent Bernoulli processes with probability $\phi_{j,i}$ for each feature i generated by latent cause j, $p(O_{i,t+1}|L_{t+1} = j) = \phi_{j,i}(t+1)$. 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 $b_{obs} > 0$. Given that pilot data from the task showed that absent any data, participants' prior expections of a US before the start of the experiment are $\sim 50\%$, we assumed symmetric Beta priors ($a_{obs} = b_{obs}$) whose mean is always 50%. Values of $a_{obs} > 1$ imply initial $\phi_{j,i}$ that are unimodal around 0.5, with higher a_{obs} being more stochastic (that is, more tightly concentrated around $\phi_{j,i} = 0.5$; Fig. 4A), while smaller a_{obs} result in bimodal $\phi_{j,i}$ that are more deterministic (i.e., closer to 0 or 1; Fig. 4C). After every observation, $\phi_{j,i}(t)$ are updated according to observed events weighted by the probability that the events were generated by latent cause j (see eq. 9 below).

Models

The basic model. Here, $N_{j,t}$ is the number of trials caused by latent cause j up to the trial t. Since there is a nonzero probability that each of the previous latent causes has generated the current trial, in practice $N_{j,t}$ sums the posterior probability of latent cause j in 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 Concentration Model" with α as the single free parameter and $a_{obs} = 1$ (i.e., a flat prior over observations given a latent cause), and a "Basic Prior Model" with a_{obs} as the single free parameter and $\alpha = 0.03$. The former model has been used previously to model fear acquisition and extinction in a similar paradigm, but not spontaneous recovery (Gershman & Hartley, 2015). The latter setting with a_{obs} as the free parameter was also used in all the models below (for a detailed explanation of the choice to use the prior rather than the concentration parameter as a free parameter, see Suppl. Subsubsec. 2.3.1).

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 (e.g., Rouhani et al., 2023), in this model we hypothesized that latent causes associated with aversive stimuli (here, the scream US) are protected from decay, perhaps due to memory reactivation (Wimmer et al., 2023). We modeled 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 scream US (O_4) given the latent cause:

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

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

Here, $p(O_4 = 1 | L_k = j)$ is given by $\phi_{j,4}$, which is updated on each trial (see the likelihood computation below, eq. 9). This model reduces to the Decay Model when $\omega = 0$.

Salience Model. Emotionally charged stimuli experienced during an event may change the salience of the event, for example, they may capture more attention (see Dolcos et al., 2020, for a review), potentially enhancing memory for the event. To account for this salience effect, we implemented a model that counts trials with an aversive US as $C \geq 0$ rather than 1 in the counts $N_{j,t}$:

$$N_{j,t} = \Sigma_{t' \leq t} P(L_{t'} = j | \mathbf{O}_{1:t}) \cdot ((1 - O_{4,t'}) + O_{4,t'} \cdot C). \tag{3}$$

Note that $O_{4,t}$ is 1 on trials with a US, and 0 otherwise. This model reduces to the basic model for $C = 1$, and ascribes extra salience to trials with aversive events when $C > 1$, but also allows for lower salience for these trials for $0 \leq C < 1$ (we will make use of this case specifically in the two models below).

Selective Maintenance and Low Salience Model. In this model, we combined the selective maintenance and salience models, restricting $C \leq 1$ (labeled C_{low} hereafter):

$$N_{j,t} = \Sigma_{t' \leq t} P(L_{t'} = j | \mathbf{O}_{1:t}) \cdot ((1 - O_{4,t'}) + O_{4,t'} \cdot C_{low}) \cdot \prod_{k=t'+1}^{t} d_{j,k}. \tag{4}$$

$$d_{j,k} = \gamma + (1-\gamma)\omega \cdot p(O_4 = 1|L_k = j). \tag{5}$$

In contrast to the selective maintenance model above, in which selective maintenance of "dangerous" latent causes occurs throughout the learning phases and breaks, this model can also capture behavior resulting from selective maintenance occurring only during breaks, because the C_{low} parameter can reduce the amount of updating for trials with aversive outcomes. In this way, low salience can counteract the effect of selective maintenance during the learning phases. However, C_{low} has no impact during breaks, where the effect of selective maintenance can fully unfold. Note that we used C_{low} rather than C to ensure good parameter recoverability, however in our exploration dataset we found similar patterns of results when relaxing this restriction. See Suppl. Subsubsec. 2.3.2 for a detailed discussion of that choice.

Decay and Low Salience Model. To compare the above model to a reduced version without selective maintenance, here we combined the Decay and Salience Model, but also restricted $C \leq 1$ (i.e., using C_{low}).

$$N_{j,t} = \Sigma_{t' \leq t} P(L_{t'} = j|\mathbf{O}_{1:t}) \cdot ((1 - O_{4,t'}) + O_{4,t'} \cdot C_{low}) \cdot \gamma^{(t-t')}. \tag{6}$$

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 $\iota \geq 0$:

$$N_{j,t} = \Sigma_{t' \leq t} P(L_{t'} = j|\mathbf{O}_{1:t}) \cdot \frac{1}{(t+1-t')^{\iota}} \tag{7}$$

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

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), in this model, for each CS, we added observations to the count $N_{j,t}$ according to a decreasing logistic function of how often that CS had been observed so far:

$$N_{j,t} = \Sigma_{t' \leq t} P(L_{t'} = j|\mathbf{O}_{1:t}) \cdot (d_{1,t'} O_{1,t'} + d_{2,t'} O_{2,t'}) \tag{8}$$

where $d_{i,t} = \frac{1}{1+e^{-\lambda(\nu + \Sigma_{t' \leq t} O_{i,t'})}}$ and $\lambda \leq 0$ and $\nu \geq 0$. Note that on each trial, O_1 or O_2 are observed and take the value 1, while the other takes the value 0 (whereas in breaks both are 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:

$$\phi_{j,i}(t) = \frac{a_{obs} + \sum_{t' \leq t} O_{i,t'} \cdot P(L_{t'} = j|\mathbf{O}_{1:t'})}{2a_{obs} + \sum_{t' \leq t} P(L_{t'} = j|\mathbf{O}_{1:t'})}. \tag{9}$$

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; Speekenbrink, 2016). Each time participants were asked to make an expectation rating, the algorithm first inferred the posterior probability over possible partitions of trials into latent causes based on all previous observations plus a current observation of either the CS+ or the CS- and used this inference to generate a prediction of the probability of the US:

$$p(O_{4,t}=1|O_{1:3,1:t},O_{4,1:t-1},\Theta,M)=\Sigma p(O_{4,t}=1|L_{1:t})\cdot p(L_{1:t}|O_{1:3,1:t},O_{4,1:t-1},\Theta,M) \tag{10}$$

where the sum was over all possible latent cause assignments approximated by averaging across the 1000 particles. This probability was used to compute the likelihoods of the participant's rating. No updating occured based on these inferences.

After observing all stimuli in a trial (CSs and USs), the posterior distribution over latent causes was calculated taking all previous observations into account. This was then used to update the observation probabilities for each latent cause $\phi_{j,i}$. Throughout, the model observed the same stimuli as the participants. To simulate breaks, we presented the model with a dummy stimulus O_3, simulating a short break between acquisition and extinction as 9 trials of O_3 and the long break between extinction and the spontaneous recovery test phase as 34 trials of O_3. These numbers of trials were chosen as the approximate number of task trials that would have occurred in the breaks in the experiment, as per the average time it took participants to complete the task trials and the break activities.

Model fitting

We fit the free parameters of each of the models above (Table 1) to the behavioral data of each 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. We computed the standard deviation σ of this normal distribution empirically for each individual as

$$\sigma=\sqrt{(}\sum_{i=1}^{2}\sum_{j=1}^{n}(g_{i,j}-p(O_{4,j}=1|O_i))^2/(n-1)) \tag{11}$$

where n is the number of rating data points (with $g_{i,j}$ being the actual rating (guess) number j for stimulus i between 0-100 divided by 100) and $p(O_{4,j}=1|O_i)$ was calculated according to equation 10 for each stimulus O_i for $i\in\{1,2\}$ and each rating timepoint j.

We computed the likelihood of each individual expectancy rating $g_{i,j}$ at timepoint t (that is, after trial $t-1$) for stimulus $i\in\{1,2\}$ as

$$p(g_{i,j}|O_{1:3,1:t},O_{4,1:t-1},\Theta,M)=0.01\cdot\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(g_{i,j}-p(O_{4,t}=1|O_{1:3,1:t},O_{4,1:t-1},\Theta,M))^2}{2\sigma^2}}. \tag{12}$$

Because participants could not register an expectation below 0 or above 100, we added the probability density for ratings below 0 and above 1 to expectancy ratings of 0 and 1, respectively. Alternative formulations of the response function, such as assuming the response error was distributed according to a beta distribution or fixing or fitting σ, led to very similar fits and the same overall pattern of results.

We used the Bayesian Adaptive Direct Search (BADS; Acerbi & Ma, 2017) algorithm for parameter search, with 24 restarts for each participant and model. To constrain the search space for the BADS algorithm, we used hard and plausible parameter bounds as specified in Table 1. Importantly, the model-fitting algorithm only received participants' expectation ratings, and no other information about the participant (such as their subgroup categorization).

Table 1: Model parameters and bounds for the fitting algorithm.

Model	**Free parameters Θ**	**Hard bounds**	**Plausible bounds**
Basic α Model	α	[0,10]	[1e-6,1]
Basic Prior Model	a_{obs}	[1e-10,10]	[1e-6,5]
Decay Model	a_{obs}	[1e-10,10]	[1e-6,5]
	γ	[0,1]	[0.5,1]
Selective Maintenance Model	a_{obs}	[1e-10,10]	[1e-6,5]
	γ	[0,1]	[0.5,1]
	ω	[0,1]	[0,1]
Salience Model	a_{obs}	[1e-10,10]	[1e-6,5]
	C	[0,10]	[0.7,2]
Selective Maintenance & Low Salience Model	a_{obs}	[1e-10,10]	[1e-6,5]
	γ	[0,1]	[0.5,1]
	ω	[0,1]	[0,1]
	C	[0,1]	[0.7,1]
Decay & Low Salience Model	a_{obs}	[1e-10,10]	[1e-6,5]
	γ	[0,1]	[0.5,1]
	C	[0,1]	[0.7,1]
Processing Loss Model	a_{obs}	[1e-10,10]	[1e-6,5]
	λ	[-10,-1e-10]	[-10, -1e-10]
	ν	[0,100]	[0 100]
Temporal Weighting Model	a_{obs}	[1e-10,10]	[1e-6,5]
	ι	[0,10]	[1e-6,5]

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 as the data were not normally distributed and reported rank-biserial correlation r as effect size for these tests. Importantly, no information about the subgroup categorization was known by the model fitting algorithm.

Statistical analyses

As parameter estimates and data ratings were bounded, they could not follow a normal distribution. Hence, we used non-parametric significance tests for comparisons between subgroups. For all these tests, we report rank-biserial correlation r as effect size.

Confirmatory analyses

We repeated the model-free analyses of the expectancy ratings (described above) on the simulated data from the winning model. We added response noise to the simulated data by replacing each simulated data point with a data point drawn from a normal distribution with the data point as μ and the empirically computed σ. Ratings below 0 and above 1 were assigned to be 0 or 1, respectively.

Preregistered replication

We preregistered the analyses for the second half of the dataset from study 1 to examine whether we can replicate the key results in an unobserved dataset. The detailed analyses plan was preregistered on OSF and can be found here:

`https://doi.org/10.17605/OSF.IO/UCZNF`.

Note that the preregistered study procedures and model fitting approach were identical to the Methods described above. Minor deviations from the analyses plan are listed in Suppl. Subsec. 1.5.

Participants

$N = 438$ participants were included in the replication dataset. After preregistering and accessing the dataset, we excluded 29 participants due to incomplete datasets and 54 participants who failed more than 2 audio attention checks. The final replication dataset contained $N = 355$ participants. $N = 172$ fulfilled the criteria for the SR group, $N = 110$ for the No SR group, $N = 56$ for the Return to Start group and $N = 11$ for the Generalizers group.

Hypothesis and statistical approach for replication

The goal of this replication was to test whether we can confirm our result that selective maintenance is the mechanism that best explains differential spontaneous recovery. To confirm this result, we required passing two tests: 1) The best model containing selective maintenance had to fit the data of the SR group better than the alternative models that did not contain selective maintenance specified in the pre-registration as per signed-rank tests for non-nested models and likelihood ratio tests for nested models; 2) The effect of selective maintenance on retention of latent cause counts, measured as $\omega' = \frac{(1-\gamma)\cdot\omega}{(1-\gamma)\cdot\omega+\gamma}$, had to be significantly larger in the SR group as compared to the No

SR group. Before applying these tests, we performed the same model fitting procedure and group assignments as detailed above and in the pre-registered analysis plan.

Replication in separate sample - Study 2

Participants and procedure

Participants were recruited to take part in an online study of a self-help tool for symptoms of depression/low mood and lack of energy. The study was approved by the Institutional Review Board of Princeton University (Protocol 15118). The goal of that separate study (clinical trial NCT06631183 on clinicaltrials.gov) was to identify predictors of improvement of depression symptoms given different types of self-help interventions (cognitive or behavioral). Participants were recruited via ResearchMatch (https://www.researchmatch.org/), social media platforms, and other methods. If they successfully passed a screener for inclusion and exclusion criteria, they were randomly assigned to a training or a test dataset with a ratio of 2:1. As part of a baseline assessment, they then completed questionnaires (including the GAD-7) and a set of behavioral tasks (including the task described in this paper). Subsequently, they engaged with a self-help tool for 5 weeks with additional symptom assessments throughout and after the time of the engagement. Similar to Study 1 in the general population, participants were not included in the analyses if they failed questionnaire or behavioral attention checks or attempted the task twice. We also excluded participants who were invited to participate in a brief zoom identity check but did not sign up or failed the identity check. See Suppl. Subsec. 1.3 for details on recruitment methods, inclusion and exclusion criteria and study design.

We had datasets from $N = 832$ participants who passed the attention checks outside of this task. We excluded $N = 65$ participants without complete datasets and $N = 247$ participants who failed audio attention checks within this task. Thus, the final symptomatic dataset contained $N = 520$ participants. Of those, $N = 209$ belonged to the SR group, $N = 123$ to the No SR group, $N = 144$ to the Return to Start group and $N = 18$ to the Generalizers group.

Assessments and statistical approach – Study 2

Follow-up data for the clinical trial are still being collected and the main analyses of the training dataset of the self-help study are still ongoing. Here, we report only on the baseline data of the fear conditioning task described above from the training dataset ($N = 520$) and not on the clinical intervention. The fear conditioning task was identical to the task in Study 1 with the exception that one more expectancy rating was placed at the end of extinction. We applied the same models described above to this symptomatic sample. We also analyzed the GAD-7 score of participants, the only anxiety measure included in that study. We used age, self-reported sex and cognitive ability (measured through a Matrix reasoning task; Zorowitz, Chierchia, Blakemore, & Daw, 2024; Chierchia et al., 2019) as control variables of no interest.

Associating symptoms of anxiety with parameter estimates

Datasets

As previous work has shown small associations between parameter estimates and symptoms in the general population (if these are found at all, e.g., Pike & Robinson, 2022), we combined the data from the exploratory and the preregistered replication dataset from Study 1 to increase statistical power for these analyses. However, we analyzed the data from the symptomatic sample in Study 2 separately given its size and the different distribution of symptom scores in that sample.

Missing data

In the complete dataset from Study 1 ($N = 671$), two participants had missing GAD-7 scores and 54 participants had missing HiTOP anxiety scores, of which $N = 46$ were in the SR or No SR group. This was due to the fact that these data were collected in a follow-up study a few weeks after the main task data were collected and not all participants returned to complete the follow-up. Participants for whom we did not have anxiety scores were excluded from corresponding analyses. In the dataset from Study 2 ($N = 520$), 4 participants had missing age, sex or cognitive ability scores and were excluded from analyses that controlled for these variables.

Statistical approach

First, we correlated symptom scores with parameter estimates. Given that we examined 7 correlations across the two datasets, we applied FDR correction to control for multiple comparison. For correlations that remained significant, we ran additional linear regressions with anxiety score as independent variable and the parameter estimate of interest, as well as parameters of no interest and covariates of age, self-reported sex and general cognitive ability as dependent variables.

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

Bouton, M. E. (2004, January). Context and Behavioral Processes in Extinction. *Learning & Memory*, *11*(5), 485–494. Retrieved 2022-10-27, from `http://learnmem.cshlp.org/content/11/5/485` (Company: Cold Spring Harbor Laboratory Press Distributor: Cold Spring Harbor Laboratory Press Institution: Cold Spring Harbor Laboratory Press Label: Cold Spring Harbor Laboratory Press Publisher: Cold Spring Harbor Lab) doi: 10.1101/lm.78804

Brown, V. M., Price, R., & Dombrovski, A. Y. (2023). Anxiety as a disorder of uncertainty: Implications for understanding maladaptive anxiety, anxious avoidance, and exposure therapy. *Cognitive, Affective, & Behavioral Neuroscience*, *23*(3), 844–868.

Chierchia, G., Fuhrmann, D., Knoll, L. J., Pi-Sunyer, B. P., Sakhardande, A. L., & Blakemore, S.-J. (2019). The matrix reasoning item bank (mars-ib): Novel, open-access abstract reasoning items for adolescents and adults. *Royal Society open science*, *6*(10), 190232.

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., Hermans, D., & Vansteenwegen, D. (Eds.). (2006). *Fear and learning: From basic processes to clinical implications*. Washington, DC, US: American Psychological Association. (Pages: xiii, 319) doi: 10.1037/11474-000

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

Craske, M. G., Treanor, M., Zbozinek, T. D., & Vervliet, B. (2022). Optimizing exposure therapy with an inhibitory retrieval approach and the optex nexus. *Behaviour Research and Therapy*, *152*, 104069.

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.

Donnellan, M. B., Oswald, F. L., Baird, B. M., & Lucas, R. E. (2006). The mini-ipip scales: tiny-yet-effective measures of the big five factors of personality. *Psychological assessment*, *18*(2), 192.

DuBrow, S., Rouhani, N., Niv, Y., & Norman, K. A. (2017). Does mental context drift or shift? *Current opinion in behavioral sciences*, *17*, 141–146.

Erev, I., & Roth, A. E. (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria. *American economic review*, 848–881.

Gagne, C., Dayan, P., & Bishop, S. J. (2018). When planning to survive goes wrong: predicting the future and replaying the past in anxiety and ptsd. *Current Opinion in Behavioral Sciences*, *24*, 89–95.

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., & Hartley, C. A. (2015, September). Individual differences in learning predict the return of fear. *Learning & Behavior*, *43*(3), 243–250. doi: 10.3758/s13420-015-0176-z

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.

Haberlandt, K., Hamsher, K., & Kennedy, A. W. (1978). Spontaneous recovery in rabbit eyelid conditioning. *The Journal of General Psychology*, *98*(2), 241–244.

Heimberg, R. G., Horner, K., Juster, H., Safren, S., Brown, E., Schneier, F., & Liebowitz, M. (1999). Psychometric properties of the liebowitz social anxiety scale. *Psychological medicine*, *29*(1), 199–212.

Hopkins, D. J., & King, G. (2010). Improving anchoring vignettes: Designing surveys to correct interpersonal incomparability. *Public opinion quarterly*, *74*(2), 201–222.

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

Löwe, B., Decker, O., Müller, S., Brähler, E., Schellberg, D., Herzog, W., & Herzberg, P. Y. (2008). Validation and standardization of the generalized anxiety disorder screener (gad-7) in the general population. *Medical care*, *46*(3), 266–274.

Mattick, R. P., & Clarke, J. C. (1998). Development and validation of measures of social phobia scrutiny fear and social interaction anxiety. *Behaviour research and therapy*, *36*(4), 455–470.

McCullough, J. P., Schramm, E., & Penberthy, J. K. (2014). *Cbasp as a distinctive treatment for persistent depressive disorder: Distinctive features*. Routledge.

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.

Pike, A. C., & Robinson, O. J. (2022, April). Reinforcement Learning in Patients With Mood and Anxiety Disorders vs Control Individuals: A Systematic Review and Meta-analysis. *JAMA Psychiatry*, *79*(4), 313. Retrieved 2023-02-16, from `https://jamanetwork.com/journals/jamapsychiatry/fullarticle/2789693` doi: 10.1001/jamapsychiatry.2022.0051

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

Rescorla, R. A., & Wagner, A. R. (1972, January). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In *Classical Conditioning II: Current Research and Theory* (Vol. Vol. 2). (Journal Abbreviation: Classical Conditioning II: Current Research and Theory)

Robbins, S. J. (1990). Mechanisms underlying spontaneous recovery in autoshaping. *Journal of Experimental Psychology: Animal Behavior Processes*, *16*(3), 235.

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

Speekenbrink, M. (2016). A tutorial on particle filters. *Journal of Mathematical Psychology*, *73*, 140–152.

Spitzer, R. L., Kroenke, K., Williams, J. B., & Löwe, B. (2006). A brief measure for assessing generalized anxiety disorder: the gad-7. *Archives of internal medicine*, *166*(10), 1092–1097.

Watson, D., Forbes, M. K., Levin-Aspenson, H. F., Ruggero, C. J., Kotelnikova, Y., Khoo, S., ... Kotov, R. (2022). The development of preliminary hitop internalizing spectrum scales. *Assessment*, *29*(1), 17–33.

Wilson, R. C., & Collins, A. G. (2019). Ten simple rules for the computational modeling of behavioral data. *Elife*, *8*, e49547.

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)

Young, J. E., Klosko, J. S., & Weishaar, M. E. (2006). *Schema therapy: A practitioner's guide*. guilford press.

Zorowitz, S., Chierchia, G., Blakemore, S.-J., & Daw, N. D. (2024). An item response theory analysis of the matrix reasoning item bank (mars-ib). *Behavior research methods*, *56*(3), 1104–1122.