Individual differences in learning predict the return of fear

Samuel J. Gershman[1] · **Catherine A. Hartley**[2]

Abstract Using a laboratory analogue of learned fear (Pavlovian fear conditioning), we show that there is substantial heterogeneity across individuals in spontaneous recovery of fear following extinction training. We propose that this heterogeneity might stem from qualitative individual differences in the nature of extinction learning. Whereas some individuals tend to form a new memory during extinction, leaving their fear memory intact, others update the original threat association with new safety information, effectively unlearning the fear memory. We formalize this account in a computational model of fear learning and show that individuals who, according to the model, are more likely to form new extinction memories tend to show greater spontaneous recovery compared to individuals who appear to only update a single memory. This qualitative variation in fear and extinction learning may have important implications for understanding vulnerability and resilience to fear-related psychiatric disorders.

✉ Samuel J. Gershman
 sjgershm@mit.edu

 Catherine A. Hartley
 cah2031@med.cornell.edu

[1] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA

[2] Sackler Institute for Developmental Psychobiology, Weill Cornell Medical College, 1300 York Avenue, Box 140, New York, NY 10065, USA

Keywords Fear conditioning · Extinction · Memory · Spontaneous recovery · Reinforcement learning

Millions of adults suffer from anxiety disorders, experiencing intense and persistent fear in their daily lives (Kessler, Chiu, Demler, & Walters, 2005). Anxiety disorders are commonly treated using therapies based on the principles of extinction learning (Hofmann, 2008). However, even after treatment, spontaneous recovery of fear often occurs (Rachman, 1989). The success or failure of treatment may hinge on individual differences in underlying learning mechanisms, but these differences remain poorly understood. Why do some people show spontaneous recovery of fear following treatment, while others do not?

We pursued an answer to this question using a laboratory procedure for the study of learned fear (Pavlovian fear conditioning) in humans. In this experimental paradigm (see Fig. 1a), a neutral conditioned stimulus (CSa) is paired with an unpleasant electric shock (the unconditional stimulus, or US). Anticipatory fear is measured by the difference in the skin conductance response (SCR) to CSa and a second stimulus that is never paired with shock (CSb). After the fear acquisition phase, presentation of CSa is no longer paired with the shock, typically leading to a gradual decrease in the fear response that is referred to as extinction. Spontaneous recovery is measured the next day by the increase in the anticipatory fear response relative to the end of extinction.

We will show that there is significant heterogeneity across individuals in the amount of spontaneous recovery, which we hypothesize might stem from qualitative differences in the nature of the learning that occurs during conditioning. Whereas some individuals may tend to form a new safety memory during extinction, leaving their fear memory intact, others may update the original threat association, effectively

A

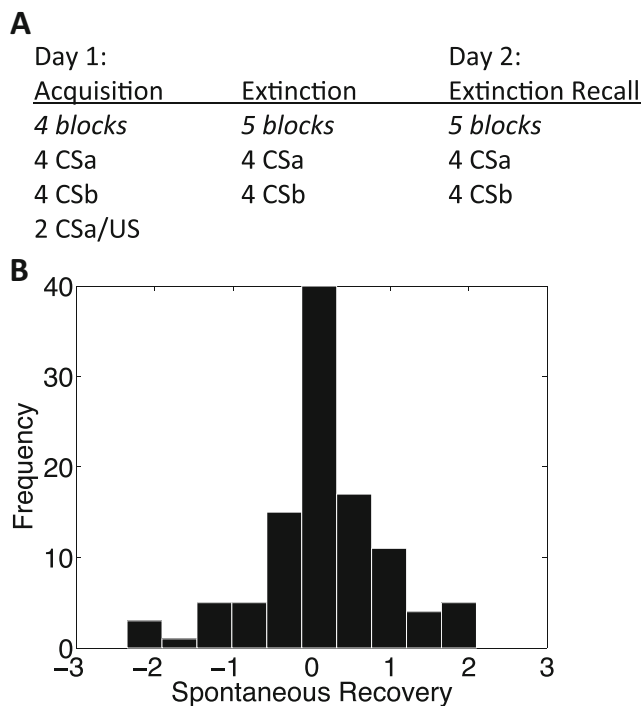| Day 1:		Day 2:
Acquisition	Extinction	Extinction Recall
4 blocks	*5 blocks*	*5 blocks*
4 CSa	4 CSa	4 CSa
4 CSb	4 CSb	4 CSb
2 CSa/US		

B

Fig. 1 (a) Experimental design. (b) Histogram of spontaneous recovery across participants. Spontaneous recovery was defined as the differential SCR on the first block of extinction recall minus the differential SCR on the last block of extinction

unlearning the fear memory. We formalize this account using a computational model of learning (Gershman, Blei, & Niv, 2010; Gershman & Niv, 2012), and show that individual differences in spontaneous recovery can be predicted from the SCR dynamics during acquisition and extinction.

The computational model posits that learners attempt to segment their experience into "states" or "latent causes" (see also Courville, Daw, & Touretzky, 2006; Gallistel, 2012; Redish, Jensen, Johnson, & Kurth-Nelson, 2007) such that each state captures a particular regularity in the configuration of observable stimuli (CS and US). A large mismatch between the inferred prototypical stimulus configuration of a state and the current stimulus configuration (akin to the prediction error in reinforcement learning models) provides evidence that a new state may be active, increasing the likelihood that it will be created (for supporting evidence, see Gershman, Jones, Norman, Monfils, & Niv, 2013; Gershman, Radulescu, Norman, & Niv, 2014). This segmentation process depends on a parameter that determines the likelihood that incongruent observations will be represented as new states. This parameter can be conceived of as modulating the amount of evidence an individual requires in order to decide that the current statistics of the environment are different from the previous state, or equivalently a threshold on the prediction error necessary to segment the recent observations into a new state. When this parameter is high, differences in CS–US stimulus configurations observed across trials will result in these trials being

assigned to distinct clusters. Intuitively, CSa trials in the acquisition phase (in which the CS is paired with the US) might be clustered together into one state, associated with a high probability of US occurrence, whereas the unreinforced CSa trials in the extinction phase might be clustered into a separate state in which expectations of the US are low. The stimulus configuration associated with each state is stored in a distinct memory, providing a computational formalization of the idea that extinction does not result in the unlearning of the "fear" memory but rather the formation of a competing "no-fear" memory (Bouton, 2004; Pavlov, 1927). According to this logic, attenuation of fear during extinction occurs because the no-fear memory temporarily inhibits the fear memory; once the inhibition wanes, spontaneous recovery can occur.

In our experiment, CSa was partially reinforced, with only one third of CSa presentations during the acquisition phase being paired with shock; thus, two thirds of CSa trials during acquisition were operationally identical to extinction trials. From the point of view of the latent cause model, this relatively high degree of similarity between acquisition and extinction trials suggests another plausible clustering in which CSa trials during both acquisition and extinction were generated by a single state. The estimated probability of US occurrence in this single state would decrease over the course of extinction, representing an alteration of the original fear memory. The latent cause model predicts the SCR on each trial as a function of the predicted probability that a US will occur, given the current inferred state. In the model, whether an individual infers a one-state or two-state clustering depends on a single parameter that specifies the individual's a priori beliefs about the structural complexity of the current environment (i.e., how many states exist). By fitting the value of this parameter that best explains the dynamics of each individual's anticipatory SCR responses, we can determine whether that person appears to have segmented reinforced and unreinforced CSa observations into distinct states or merged them into a single state in which threat subsides during the course of extinction.

The latent cause model provides insight into why some people might show spontaneous recovery while others do not. In essence, subjects who showed spontaneous recovery learned to store the acquisition trials in a separate memory from the extinction trials, allowing the fear memory to be retrieved later. In contrast, subjects who showed no spontaneous recovery learned to store all trials in the same memory, thereby effectively unlearning the fear memory during extinction. Given our presently incomplete understanding of how fear and extinction memories are stored in the brain (see Pape & Pare, 2010, for a review), this interpretation is necessarily speculative and comes with a number of caveats (e.g., other assays of fear memory, such as renewal or reinstatement, might have revealed the presence of an intact fear memory). Nonetheless, our results cannot be explained solely in terms of the longstanding assumption that extinction represents new

learning (Bouton, 2004; Pavlov, 1927). In moving beyond the behavior of the group and examining individual variation in learning, our results are consistent with the idea that for a large proportion of individuals, extinction might reflect unlearning.

Materials and methods

The experimental data set we analyzed was previously published by Hartley et al. (2012), and we refer readers to that paper for detailed experimental methods. Note that because this is a reanalysis of existing data, the sample size was determined independently of the analyses we present below. Below, we summarize the experimental methods and then provide a brief description of the probabilistic model. More details about the probabilistic model can be found in the supplemental information (SI) text and in Gershman and Niv (2012).

Participants One hundred forty-one volunteers, ages 18–35 ($M = 21.1$, $SD = 3.5$, 87 female) were recruited at New York University. All participants gave informed consent and were paid for participation. Participants were not taking any psychiatric medication. Twelve participants were excluded because of experimental error, data corruption, or a failure to display a variable skin conductance signal. Another 23 participants were excluded because of failure to exhibit a discriminative conditioned response during the end of the acquisition phase, leaving 106 participants whose data we analyzed in this paper.

Procedure Participants underwent a 2-day fear conditioning procedure. The first day consisted of two phases: an acquisition phase followed immediately by an extinction phase. On each trial in the acquisition phase, one of two colored squares was presented. One square (CSa) coterminated with a mild electric wrist shock on 33 % of trials, whereas the other square (CSb) was never paired with shock. Trials in the acquisition phase were organized into four blocks (see Fig. 1a), each consisting of two reinforced CSa trials, four unreinforced CSa trials, and four CSb trials. Trials in the extinction phase were organized into five blocks, each consisting of four unreinforced CSa trials and four CSb trials. On the second day (extinction recall), participants returned for five additional blocks of extinction.

The SCR was recorded through shielded Ag-AgCl electrodes attached to the second and third fingers of the left hand. SCR data were low-pass filtered and smoothed. The greatest base to peak change in SCR in a 0.5- to 4.5-s window after each CS onset was assessed for each trial. Only unreinforced CSa trials are included in all our analyses, since the slow timescale of the SCR makes it difficult to distinguish anticipatory responses from reaction to the US. Our analyses focus on the differential SCR, which is obtained for each block by subtracting the mean response to CSb from the mean response to CSa. The differential SCR thus reflects the degree to which learned fear exceeds the baseline response to a stimulus unassociated with shock. Our measure of spontaneous recovery was the increase in this differential SCR response from the final block of day-1 extinction to the initial block of extinction recall: (Extinction Recall, Block 1: Mean CSa – Mean CSb) - (Extinction, Block 5: Mean CSa – Mean CSb).

Computational model We fit the computational model described in Gershman and Niv (2012) to each participant's data separately. We used the raw SCR response as input to the model fitting, although for statistical analysis and visualization we report the differential SCR. The model was fit to the acquisition blocks and the first four blocks of extinction; it is therefore unbiased by the level of spontaneous recovery.

Here we present the model informally; see the SI text for more technical details. The model assumes that on each trial, participants compute, using Bayes' rule, the posterior probability that state s generated the observed stimuli:

$$P\left(state = s \mid stimuli\right) \propto P\left(stimuli \mid state = s\right) P(state = s) \quad (1)$$

Each state is associated with a predicted stimulus configuration—roughly, the average stimulus configuration of all trials assigned to a particular state. The likelihood $P(stimuli|state = s)$ expresses the consistency between the current stimuli and the predicted configuration associated with hypothetical state s. The prior $P(state = s)$ expresses a learner's preference for "simpler" clusterings (i.e., with a small number of states). Specifically, the prior biases the model to assign new trials to a given state in proportion to the number of previous trials assigned to the state; with probability proportional to α (the only parameter governing the prior), a trial will be assigned to a new state. Smaller values of α thus induce a stronger tendency to form simple clusterings (i.e., clusterings that assign observations to a small number of states). As α grows, the preference for simplicity diminishes, and the clustering eventually assigns each trial a unique state.

While Bayes' rule stipulates the optimal probabilistic computation, this computation is not in general tractable, because to normalize the probability distribution requires summing over all possible clusterings (which grows exponentially with the number of trials). For this reason, a practical implementation of the model requires approximating Bayesian inference. We do not wish to make any strong claims about the precise approximation that animals may be using because this is underconstrained by the data, but we describe two tractable and psychologically plausible approximations in the SI text. Both of these approximations rely on summing over a small number of high probability clusterings. It has been suggested that this kind of "sampling" approximation can explain why

individual learning curves sometimes appear abrupt and unstable (Daw & Courville, 2008).

The model generates a prediction of the SCR on each trial as a linear function of the predicted probability that a US will occur:

$$SCR = \beta \sum_s P\left(US \middle| state = s\right) P\left(state = s \middle| stimuli\right) + \varepsilon \quad (2)$$

where β is a scaling parameter that maps US probability to SCR and ϵ is a noise parameter drawn from a normal distribution with mean = 0 and variance = 1. Larger values of β mean that the SCR is more sensitive to changes in US probability.

Details of how α and β were fit to participants' data can be found in the SI text. Briefly, we used nonlinear optimization to find the values of the parameters that maximized the likelihood of the SCR data.

Results

Spontaneous recovery of fear is heterogeneous We measured the spontaneous recovery of fear in 106 participants (see Materials and Methods) by calculating the increase in differential SCR from the final block of extinction on day 1 to the first block of extinction recall on day 2: (Extinction Recall, Block 1: Mean CSa – Mean CSb) - (Extinction, Block 5: Mean CSa – Mean CSb). We found that there was substantial heterogeneity across the group (see Fig. 1b), with 43 participants showing *negative* spontaneous recovery (i.e., the differential SCR to CSa vs. CSb during the first block of recall was smaller than during the last block of extinction). Overall, the 95 % confidence interval was +/- 0.16 around a mean spontaneous recovery of 0.14. We could not reject the null hypothesis that the mean was equal to 0 ($p = .09$, one-sample t test).

A computational model reveals two groups of participants Although Fig. 1b suggests a unimodal distribution centered near 0, this analysis conceals hidden structure in our data. We obtained a finer-grained characterization of responses during the acquisition and extinction phases by fitting a computational model of learning (the latent cause model; Gershman et al., 2010; Gershman & Niv 2012) to each participant's data. Importantly, we fit the model to data from the acquisition phase and the first four blocks of extinction, thus ensuring that it was unbiased by spontaneous recovery performance.

According to the latent cause model, participants assign each trial to a latent cause or state; the collection of assignments constitutes a clustering of trials (see Materials and Methods and SI text for more details). We found that participants fell into two groups: a "one-state" group ($n = 88$) that assigned all trials to the same state and a "two-state" group ($n = 18$) that assigned trials primarily to either of two states (in some cases more than two states were invoked, but these additional states had low probability). The division of participants into the two groups can be seen clearly in Fig. 2a, which shows the marginal probability of each state across participants. The marginal probability of a third state is never greater than .15, with a median probability of .03.

The two groups exhibited strikingly different learning dynamics (Fig. 2b), with the one-state group acquiring fear more slowly than the two-state group. The latent cause model posits that participants in the one-state group learn more slowly because they tend to cluster CSa and CSa trials together, whereas the two-state group tends to separate them into different clusters.

Despite the fact that the two groups showed no difference in fear on the last block of extinction ($p = .56$, two-sample t test), participants in the two-state group showed significantly higher recovery of fear on the first block of recall compared to participants in the one-state group, $t(104) = 3.53$, $p < .001$, two-sample t test; see Fig. 2. We can quantify the change in fear from extinction to recall using a spontaneous recovery score (see Fig. 3a), measured as the difference in SCR between the first block of recall and the last block of extinction. The spontaneous recovery score for participants in the one-state group was not significantly different from 0 ($p = .79$, one-sample t test), in contrast to participants in the two-state group, who showed spontaneous recovery significantly greater than 0, $t(17) = 4.43$ $p < .0001$, one-sample t test. Furthermore, the two groups were significantly different from each other, $t(104) = 3.36$, $p < .005$, two-sample t test. The same result was obtained when only looking at the SCR to CSa, $t(104) = 3.99$, $p < .001$, two-sample t test; see SI for separate plots of CSa and CSa responses.

Rather than doing a hard split of participants, we can also look at a continuous measure of each participant's tendency to cluster their experience into one or two states. This tendency is captured by the parameter α, which expresses an individual's a priori preference for a "simpler" clustering with a small number of states. We fit this parameter to each participant separately. Our fitting method (see SI text) furnishes us with the log probability that $\alpha > 0$, which we refer to as the log Bayes factor (log BF) since it represents a Bayesian metric for comparing models with $\alpha > 0$ and $\alpha = 0$. The log BF captures a participant's preference for clusterings with more than one state. Figure 3b plots each participant's spontaneous recovery against the participant's log BF, showing that the two measures are positively correlated ($r = .40$, $p < .0001$). Thus, participants who exhibit a stronger preference for clusterings with more than one state tend to show greater spontaneous recovery.

Spontaneous recovery is well explained by the model, but not by other individual differences We next investigated how well the grouping of participants derived from our model explains spontaneous recovery when controlling for
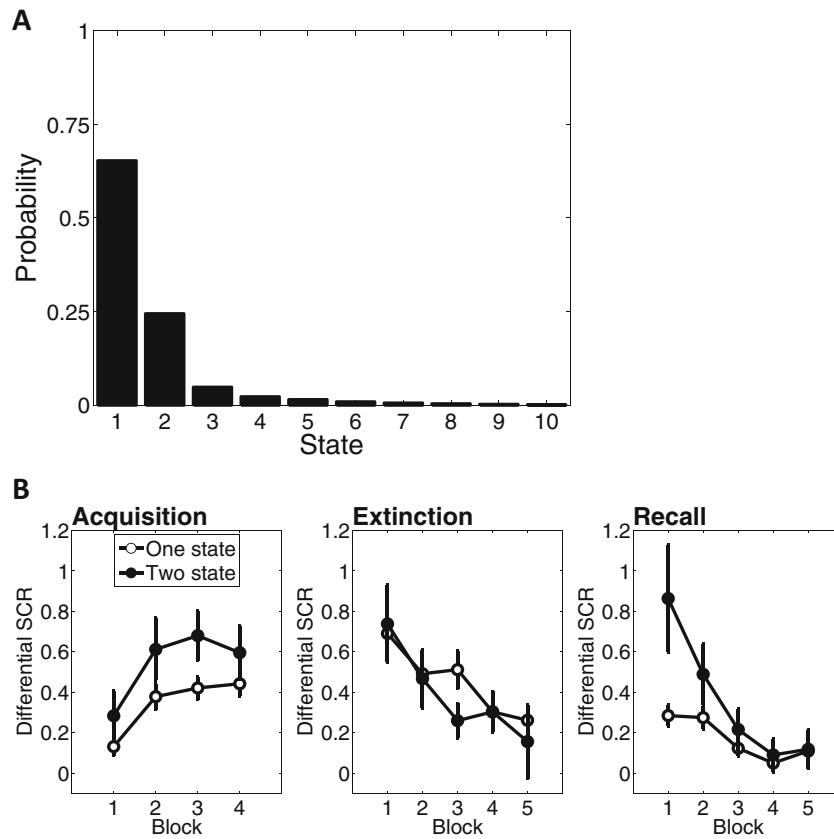
Fig. 2 (a) Marginal probability of each state being active, averaged across participants. (b) Differential SCR over the course of acquisition, extinction, and extinction recall, shown separately for the one-state and two-state groups. Error bars show standard error of the mean

various other individual differences. We ran a multiple linear regression with spontaneous recovery as the response variable and six standardized covariates: state and trait anxiety on the State-Trait Anxiety Inventory, age, number of serotonin transporter polyadenylation polymorphism (STPP/rs3813034) G alleles, sex, and the log BF. A greater number of STPP G alleles was previously shown to be weakly predictive of spontaneous recovery in this data set (Hartley et al., 2012). Figure 3c shows the estimated regression coefficient for each covariate. Only the log BF had a coefficient that was
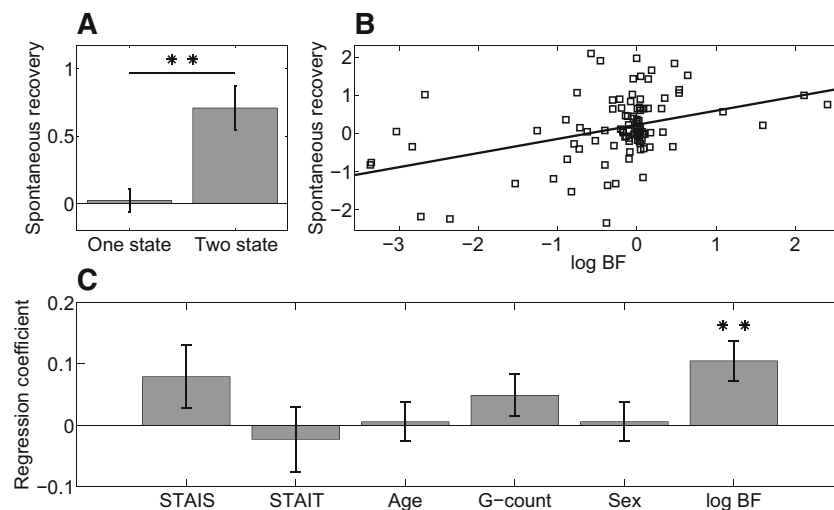

Fig. 3 (a) Spontaneous recovery for the one-state and two-state groups. (b) Each participant's spontaneous recovery plotted against the log probability that $\alpha > 0$ (log Bayes factor). (c) Regression coefficients for standardized covariates: state anxiety (STAIS), trait anxiety (STAIT), age, number of STPP G alleles (G-count), sex, and log Bayes factor. Error bars show standard error of the mean. ** = $p < .01$, Bonferroni corrected

significantly greater than 0 ($p < .01$, one-sample t test, Bonferroni corrected). We conclude that among the covariates we considered, only the model-derived group assignment was a robust predictor of spontaneous recovery.

We also examined other individual differences within the learning and extinction data. The asymptotic level of learning (i.e., the differential SCR on the final block of conditioning) did not predict spontaneous recovery ($r = .01$, $p = .89$; Supplementary Figure 2). The learning rate, which we estimated in a model-free manner by fitting an exponential curve to the learning data, also did not predict spontaneous recovery ($r = .002$, $p = .99$). The degree of freezing on the eighth block of extinction did significantly predict spontaneous recovery ($r = -.48$, $p < .0001$). However, we explicitly controlled for this factor in our model-based clustering by separating participants into two groups using a cluster chosen to minimize the difference between the average differential SCR for the two groups on the eighth block of extinction (see supplemental information). Accordingly, the one-state and two-state groups did not differ significantly in differential SCR on the eighth block of extinction ($p = .99$, two-sample t test). Controlling for the degree of freezing on the eighth block of extinction, there was still a significant partial correlation between log BF and spontaneous recovery ($r = .29$, $p < .005$). Thus, there is a significant amount of variation in our spontaneous recovery data that is not explained by simple summary statistics of the learning and extinction data.

If the model is not using these individual statistics to discriminate between the two groups, what is it using? Each of these statistics captures differences in learning about the changing configurations of CS and US stimulus presentations over time. The model is in fact using a combination of these statistics, combining them in a computationally principled way to estimate differences in how individuals have clustered stimulus observations into states. The learning and extinction rates, as well as their asymptotes, all provide individually weak signals, but the model integrates these weak signals to identify a reliable distinction between sub-groups of participants.

More specifically, the model discriminates between the two groups by jointly detecting the following learning characteristics. First, the one-state group learns more slowly because the reinforced CSa and the unreinforced CSb are both assigned to the same state, thereby reducing the effective reinforcement accruing to the single state. In contrast, the two state group assigns CSa and CSb to different states and hence can acquire a conditioned response to CSa more quickly. Second, the two-state group extinguishes its response to CSa more quickly (see Supplementary Figure 1) because it infers that a new state is active in the extinction phase, whereas the one-state group explains both phases with a single state. This is essentially a version of the partial reinforcement extinction effect (see Gershman & Niv, 2012, for more discussion).

Looking inside the model One reason to use a model is that we can look at its internal variables to better understand the cognitive processes giving rise to behavior. Figure 4 shows the evolution of a simulated participant's probability distribution over states for unreinforced CSa trials, given three different values of the α parameter. We can see that the posterior probability of a new state increases over the course of acquisition, with a more gradual increase for smaller values of α. This increase occurs because the partial reinforcement schedule provides evidence that reinforced and unreinforced CSa trials are generated by different states, opposing the model's a priori preference for a one-state clustering, and hence the model becomes increasingly ambivalent between the one- and two-state clustering. During extinction, the lack of reinforcement tips the balance in favor of a two-state clustering, but only for participants with a sufficiently large value of α; for participants with a small value of α, the probability of a new state always remains small.

If we equate states with memories, then the foregoing explanation resonates strongly with the popular view that extinction involves the formation of a new memory (Bouton, 2004; Pavlov, 1927). The latent cause model goes further, explaining individual differences in probabilistic terms: a stronger preference for simpler clusterings leads to a higher likelihood of encoding both acquisition and extinction into a single memory.

Comparison with the rescorla–Wagner model Can other computational models explain individual differences in spontaneous recovery? We addressed this question by fitting the renowned Rescorla–Wagner model (Rescorla & Wagner, 1972) to the acquisition and extinction data. This model, which has the same number of free parameters as the latent cause model (see Materials and Methods), exhibited a poorer fit to the data, as measured by the log likelihood of the SCR: Every participant was better fit by the latent cause model. Breaking down the log likelihood ratio according to the model-based group assignments, we found that the log likelihood ratio was significantly smaller for the one-state group
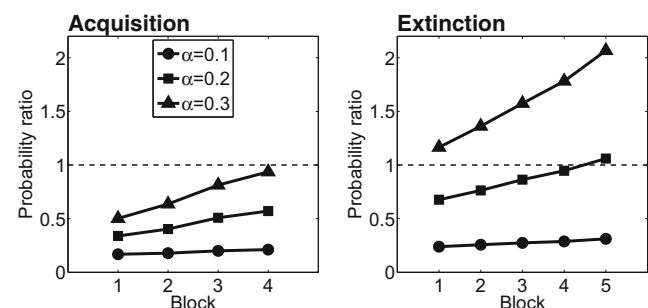

Fig. 4 Simulations of state inference using different values of the α parameter. The y-axis shows the ratio p(new state)/p(old state). Values above 1 (indicated by a dashed line) represent blocks on which the probability of a new state exceeds the probability of an old state

compared to the two-state group—$t(104) = 8.79$, $p < .00001$, two-sample t test—indicating stronger support for the latent cause model in those participants who are best described by multiple states (see Fig. 5). Thus, the Rescorla–Wagner model can provide a reasonable fit for a subgroup of participants but was unable to explain the larger pattern of individual differences, ostensibly because it makes no provision for the formation of new states.

Discussion

In a study of fear conditioning, we found substantial heterogeneity across participants in the degree of spontaneous recovery following extinction. At the group level, the spontaneous recovery of extinguished fear is a well-documented phenomenon (Pavlov, 1927; Rescorla, 2004). However, in striking contrast to this average behavior, approximately 83 % of our participants showed no evidence of spontaneous recovery. A growing literature suggesting that the spontaneous recovery of fear may confer heightened risk of anxiety underscores the need to clarify the mechanisms underlying this marked heterogeneity in fear learning and extinction (Hartley & Casey, 2013; Milad & Quirk, 2012).

Consistent with our theoretical account, we found that participants could be divided into two groups: one group that preferred assigning all acquisition and extinction trials to a single state, and one group that preferred assigning acquisition trials to one state and extinction trials to another state. Participants in the one-state group showed no evidence of spontaneous recovery, while participants in the two-state group showed significant spontaneous recovery. Strikingly, this grouping was found to be the only reliable predictor of spontaneous recovery among a variety of other individual differences.

These qualitative differences in fear learning may reflect distinct underlying neural processes in participants in the one-

Fig. 5 Log likelihood ratio of the latent cause model relative to the Rescorla-Wagner model, shown separately for the one-state and two-state groups. Error bars show standard error of the mean

versus two-state groups. Research in animal models suggests that distinct populations of neurons within the amygdala are active during fear acquisition and extinction (Herry et al., 2008), which compete for the context-dependent control of behavior via dynamic interaction with regions of the medial prefrontal cortex and the hippocampus (Farinelli, Deschaux, Hugues, Thevenet, & Garcia, 2006; Maren, Phan, & Liberzon, 2013; Milad & Quirk, 2012). Two-state learners may exhibit more sensitive pattern separation, a hippocampal-dependent process through which regularities in CS-outcome observations might be distinguished (Marr, 1971), facilitating the formation distinct fear and extinction memories. In contrast, one-state learners may alter the original fear memory within the amygdala, updating this representation with new safety information acquired over the course of learning. These speculative hypotheses might be tested in a future study examining the neural correlates of these two forms of learning.

One concern with the approach pursued here is that our theoretical interpretation rests upon rather complex Bayesian machinery. While the model only has a single parameter (and hence is "simple" from a statistical point of view), there are several equations jointly governing the model, making it appear more "complex" compared to some existing models of learning (e.g., Rescorla & Wagner, 1972). However, complexity in this latter sense is inherently subjective; the Bayesian formalism may be unfamiliar to many readers, but its computations are in fact tightly constrained by the rules of probability theory and assumptions about the animal's internal model of the environment (which in some cases can be verified or estimated; see, for example, Stocker & Simoncelli, 2006).

Another reason to tolerate complexity is that any reasonably comprehensive model of animal learning will necessarily be complex given the multifaceted nature of learning, involving interacting processes such as attention, memory, motivation, and goals (Balleine & Dickinson, 1998; Kutlu & Schmajuk, 2012; Wagner, 1981). In this vein, it is worth noting that our model was not manufactured de novo to fit the experimental data reported in this paper. It has been applied to a wide range of animal learning phenomena (Gershman et al., 2010; Gershman & Niv, 2012), and is derived from computational principles that appear to be common across cognitive domains (Austerweil, Gershman, Tenenbaum, & Griffiths, 2015).

Individual variation in fear extinction is proposed to modulate vulnerability to anxiety disorders, as well as the efficacy of treatment (see Milad & Quirk, 2012, for a review). The marked heterogeneity in such clinical outcomes underscores the importance of understanding deviation from the "average." For example, Holmes and Singewald (2013) pointed out that although as many as 75 % of U.S. adults may be exposed to at least one severe trauma in their lifetime (Breslau & Kessler, 2001), only 7 % of adults manifest posttraumatic stress disorder (Kessler et al., 2005). Understanding such
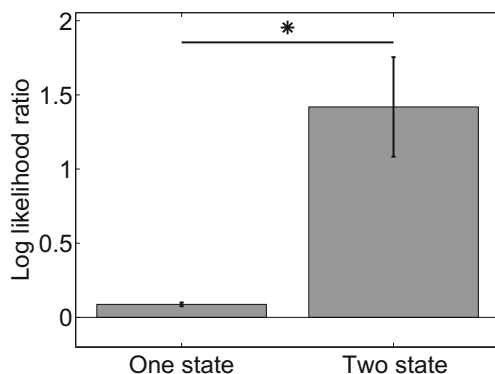
psychiatric vulnerability and resilience will require a more complete mechanistic model of the cognitive and neural heterogeneity in fear learning and extinction. New insights from structural brain imaging (Hartley, Fischl, & Phelps, 2011) and selective breeding of rats for particular extinction-related phenotypes (Bush, Sotres-Bayon, & LeDoux, 2007) have begun to articulate the neural and behavioral profiles of relevant individual differences. Our hope is that future computational theories of extinction and recovery will integrate and attempt to account for such individual differences in learning.

References

Austerweil, J. L., Gershman, S. J., Tenenbaum, J. B., & Griffiths, T. L. (2015). Structure and flexibility in Bayesian models of cognition. In J. R. Busemeyer, J. T. Townsend, Z. Wang, & A. Eidels (Eds.), *Oxford handbook of computational and mathematical psychology* (pp. 187–208). New York, NY: Oxford University Press.

Balleine, B. W., & Dickinson, A. (1998). Goal-directed instrumental action: Contingency and incentive learning and their cortical substrates. *Neuropharmacology, 37,* 407–419.

Bouton, M. E. (2004). Context and behavioral processes in extinction. *Learn Mem, 11,* 485–494.

Breslau, N., & Kessler, R. C. (2001). The stressor criterion in DSM-IV posttraumatic stress disorder: An empirical investigation. *Biol Psychiatry, 50,* 699–704.

Bush, D. E., Sotres-Bayon, F., & LeDoux, J. E. (2007). Individual differences in fear: Isolating fear reactivity and fear recovery phenotypes. *J Trauma Stress, 20,* 413–422.

Courville, A. C., Daw, N. D., & Touretzky, D. S. (2006). Bayesian theories of conditioning in a changing world. *Trends Cogn Sci, 10,* 294–300.

Daw, N., & Courville, A. (2008). The pigeon as particle filter. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in neural information processing systems* (pp. 369–376). Cambridge, MA: MIT Press.

Farinelli, M., Deschaux, O., Hugues, S., Thevenet, A., & Garcia, R. (2006). Hippocampal train stimulation modulates recall of fear extinction independently of prefrontal cortex synaptic plasticity and lesions. *Learn Mem, 13,* 329–334.

Gallistel, C. R. (2012). Extinction from a rationalist perspective. *Behav Process, 90,* 66–80.

Gershman, S. J., Blei, D. M., & Niv, Y. (2010). Context, learning and extinction. *Psychol Rev, 117,* 197–209.

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. *Front Behav Neurosci, 7*, doi:10.3389/fnbeh.2013.00164

Gershman, S. J., & Niv, Y. (2012). Exploring a latent cause model of classical conditioning. *Learn Behav, 40,* 255–268.

Gershman, S. J., Radulescu, A. R., Norman, K. A., & Niv, Y. (2014). Statistical computations underlying the dynamics of memory updating. *PLoS Comput Biol, 10,* e1003939.

Hartley, C. A., & Casey, B. J. (2013). Risk for anxiety and implications for treatment: Developmental, environmental, and genetic factors governing fear regulation. *Ann N Y Acad Sci, 1304,* 1–13.

Hartley, C. A., Fischl, B., & Phelps, E. A. (2011). Brain structure correlates of individual differences in the acquisition and inhibition of conditioned fear. *Cereb Cortex, 21,* 1954–1962.

Hartley, C. A., McKenna, M. C., Salman, R., Holmes, A., Casey, B. J., Phelps, E. A., & Glatt, C. E. (2012). Serotonin transporter polyadenylation polymporphism modulates the retention of fear extinction memory. *Proceed Nat Acad Sci, 109,* 5493–5498.

Herry, C., Ciocchi, S., Senn, V., Demmou, L., Müller, C., & Lüthi, A. (2008). Switching on and off fear by distinct neuronal circuits. *Nature, 454,* 600–606.

Hofmann, S. G. (2008). Cognitive processes during fear acquisition and extinction in animals and humans: Implications for exposure therapy of anxiety disorders. *Clin Psychol Rev, 28,* 199–210.

Holmes, A., & Singewald, N. (2013). Individual differences in recovery from traumatic fear. *Trends Neurosci, 36,* 23–31.

Kessler, R. C., Chiu, W. T., Demler, O., & Walters, E. E. (2005). Prevalence, severity, and comorbidity of twelve-month DSM-IV disorders in the National Comorbidity Survey Replication (NCSBR). *Arch Gen Psychiatry, 62,* 617–627.

Kutlu, M. G., & Schmajuk, N. A. (2012). Solving Pavlov's puzzle: Attentional, associative, and flexible configural mechanisms in classical conditioning. *Learn Behav, 40*, 269–291.

Maren, S., Phan, K. L., & Liberzon, I. (2013). The contextual brain: implications for fear conditioning, extinction and psychopathology. *Nat Rev Neurosci, 14,* 417–428.

Marr, D. (1971). Simple memory: A theory for archicortex. *Phil Trans .Royal Soc London B, Biol Sci, 262,* 23–81.

Milad, M. R., & Quirk, G. J. (2012). Fear extinction as a model for translational neuroscience: Ten years of progress. *Annu Rev Psychol, 63,* 129–151.

Pape, H. C., & Pare, D. (2010). Plastic synaptic networks of the amygdala for the acquisition, expression, and extinction of conditioned fear. *Physiol Rev, 90,* 419–463.

Pavlov, I. P. (1927). *Conditioned reflexes*. Oxford, England: Oxford University Press.

Rachman, S. (1989). The return of fear: Review and prospect. *Clin Psychol Rev, 9,* 147–168.

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

Rescorla, R. A. (2004). Spontaneous recovery. *Learn Mem, 11,* 501–509.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II* (pp. 64–69). New York, NY: Appleton-Century-Crofts.

Stocker, A. A., & Simoncelli, E. P. (2006). Noise characteristics and prior expectations in human visual speed perception. *Nat Neurosci, 9,* 578–585.

Wagner, A. R. (1981). SOP: A model of automatic memory processing in animal behavior. In N. E. Spear & R. R. Miller (Eds.), *Information processing in animals: Memory mechanisms* (pp. 5–47). London, England: Routledge.