Behavioral/Cognitive

Role of Human Ventromedial Prefrontal Cortex in Learning and Recall of Enhanced Extinction

Joseph E. Dunsmoor,[1] Marijn C.W. Kroes,[2] Jian Li,[3] Nathaniel D. Daw,[4] Helen B. Simpson,[5,6] and Elizabeth A. Phelps[7]

[1]Department of Psychiatry, University of Texas at Austin, Austin, Texas 78712, [2]Department of Cognitive Neuroscience, Donders Institute for Brain, Cognition and Behaviour, Radboud University Nijmegen Medical Centre, Nijmegen, The Netherlands, [3]School of Psychological and Cognitive Sciences and Beijing Key Laboratory of Behaviour and Mental Health, Peking University, Beijing, China, [4]Princeton Neuroscience Institute and Department of Psychology, Peking University, Princeton, New Jersey 08544, [5]Department of Psychiatry, Columbia University, New York, New York 10032, [6]New York State Psychiatric Institute, New York, New York 10032, and [7]Department of Psychology, Harvard University, Cambridge, Massachusetts 02138

Standard fear extinction relies on the ventromedial prefrontal cortex (vmPFC) to form a new memory given the omission of threat. Using fMRI in humans, we investigated whether replacing threat with novel neutral outcomes (instead of just omitting threat) facilitates extinction by engaging the vmPFC more effectively than standard extinction. Computational modeling of associability (indexing surprise strength and dynamically modulating learning rates) characterized skin conductance responses and vmPFC activity during novelty-facilitated but not standard extinction. Subjects who showed faster within-session updating of associability during novelty-facilitated extinction also expressed better extinction retention the next day, as expressed through skin conductance responses. Finally, separable patterns of connectivity between the amygdala and ventral versus dorsal mPFC characterized retrieval of novelty-facilitated versus standard extinction memories, respectively. These results indicate that replacing threat with novel outcomes stimulates vmPFC involvement on extinction trials, leading to a more durable long-term extinction memory.

Key words: extinction; fMRI; inhibitory learning; Pavlovian conditioning; ventromedial prefrontal cortex

Significance Statement

Psychiatric disorders characterized be excessive fear are a major public health concern. Popular clinical treatments, such as exposure therapy, are informed by principles of Pavlovian extinction. Thus, there is motivation to optimize extinction strategies in the laboratory so as to ultimately develop more effective clinical treatments. Here, we used functional neuroimaging in humans and found that replacing (rather than just omitting) expected aversive events with novel and neutral outcomes engages the ventromedial prefrontal cortex during extinction learning. Enhanced extinction also diminished activity in threat-related networks (e.g., the insula, thalamus) during immediate extinction and a 24 h extinction retention test. This is new evidence for how behavioral protocols designed to enhance extinction affects neurocircuitry underlying the learning and retention of extinction memories.

Introduction

Animals are exceptionally good at learning and retaining associations between environmental cues and threatening events. However, it is difficult to change these associations if threat cues are later experienced as safe. This imbalance between expressions of threat and safety is captured by Pavlovian threat (fear) conditioning, wherein conditioned defensive behaviors reemerge following extinction (Bouton, 2002). As the principles of extinction form the basis for exposure therapy (Foa and Kozak, 1986; Milad and Quirk, 2012; Vervliet et al., 2013; Ball et al., 2017), there is motivation to develop behavioral strategies that more effectively prevent the relapse of maladaptive behavior (Craske et al., 2008, 2014; Dunsmoor et al., 2015a). But it remains unclear how behaviorally enhanced extinction strategies affect neural processes underlying the learning and retention of extinction memories.

In a set of cross-species behavioral threat conditioning experiments (Dunsmoor et al., 2015c), we showed that replacing an

Received Oct. 18, 2018; revised Jan. 29, 2019; accepted Feb. 1, 2019.

Author contributions: J.E.D. and E.A.P. designed research; J.E.D. performed research; J.E.D., M.C.W.K., and J.L. analyzed data; J.E.D. wrote the first draft of the paper; J.E.D., M.C.W.K., J.L., N.D.D., H.B.S., and E.A.P. edited the paper; J.E.D., M.C.W.K., J.L., N.D.D., H.B.S., and E.A.P. wrote the paper.

This work was supported by National Institutes of Health Grant R01 MH097085 to E.A.P. and Grant K99R00 MH106719 to J.E.D. M.C.W.K. was supported by an H2020 Marie Sklodowska-Curie Fellowship and a Branco Weiss Fellowship, Society in Science. We thank Michael Evans for assistance with data collection.

The authors declare no competing financial interests.

Correspondence should be addressed to Joseph E. Dunsmoor at joseph.dunsmoor@austin.utexas.edu.

https://doi.org/10.1523/JNEUROSCI.2713-18.2019

expected aversive outcome (electric shock) with a novel, surprising, and neutral outcome (a tone) on extinction trials enhanced the long-term effects of extinction training. That is, compared with merely omitting shocks, replacing shocks with a neutral outcome (a procedure we referred to as novelty-facilitated extinction [NFE]) improved 24 h retention of extinction, evidence by diminished conditioned skin conductance responses (SCRs) in humans and freezing in rats to the conditioned stimulus (CS) (see also Lucas et al., 2018). One potential mechanism by which replacing threat with novel outcomes improved extinction retention concerns the general role played by surprise in associative learning (Rescorla and Wagner, 1972). Most associative learning models describe extinction as new learning generated by the surprising absence of an expected unconditioned stimulus (US) (Pearce and Hall, 1980; Wagner, 1981; Larrauri and Schmajuk, 2008). In popular computational learning models (Pearce and Hall, 1980), surprise generated by the omission of the US governs the rate and effectiveness of extinction by modulating a property of the CS referred to as "associability." A core feature of associability is that the ability for the CS to enter into a new association is dynamically determined by the unsigned (absolute value) prediction error on the previous trial. In this way, an associability model would predict that maximizing surprise on extinction trials could accelerate and strengthen a secondary (nonthreat) association.

The generation of a secondary competing association during extinction is linked to interactions between the amygdala and ventromedial prefrontal cortex (vmPFC), a region necessary for the formation of extinction memories (Myers and Davis, 2002; Quirk and Mueller, 2008; Hartley and Phelps, 2010; Giustino and Maren, 2015). Electrical (Milad et al., 2004) and optogenetic (Do-Monte et al., 2015) stimulation of the vmPFC on extinction trials enhances learning and retention of extinction memories in rodents. Enhancing vmPFC activity immediately after extinction via dopamine administration also improves extinction retention (Haaker et al., 2013). It is possible that increasing associability to the CS is one way to (behaviorally) stimulate engagement of the vmPFC to accelerate new associative learning. Because the strength of associative memory retention is proportional to the effectiveness of learning (Miller and Laborda, 2011; Laborda and Miller, 2012), stimulating involvement of extinction-related brain regions should then lead to the formation of a more durable extinction memory to combat expression of the original threat memory at a future test (Bouton, 2002). The results of enhanced extinction might then be revealed at test via strengthened amygdala-vmPFC functional connectivity involved in balancing the expression of high or low fear states (Quirk et al., 2003; Likhtik et al., 2005; Krabbe et al., 2018).

We examined neural mechanisms by which enhanced extinction strengthens extinction learning and diminishes the return of threat. We predicted that computational modeling would characterize enhanced extinction as stimulating involvement of the vmPFC during new learning. Further, we predicted corresponding diminution of activity in canonical threat appraisal and expression regions (e.g., the insula, thalamus, and dorsal anterior cingulate cortex [dACC]) during extinction learning and retention test, and stronger connectivity between the amygdala and vmPFC at test.

Materials and Methods

Participants. Forty-eight right-handed participants who were self-reportedly free of neurological and psychiatric disorders were scanned for this experiment. Data from 2 subjects were unusable because they did not return for the second day. The remaining 46 participants (29 female; mean age 23.43 years; SD = 4.76 years; age range 18–35 years) were randomly assigned to the standard extinction group (EXT; $N = 23$; 14 females) or NFE group ($N = 23$; 13 females). All subjects provided written informed consent approved by the University Committee on Activities Involving Human Subjects at New York University (Institutional Review Board #20162).

Task and procedure. This was a between-subjects design that included Pavlovian threat conditioning, immediate extinction, and a 24 h test, based on the human behavioral experiment described by Dunsmoor et al. (2015c) (see Fig. 1*A*). The first day included 4 functional imaging runs of equal length: threat conditioning (2 runs) and extinction (2 runs); the second day included retention test. Before scanning on the first day, the electric shock was attached to the right wrist and calibrated to be at level deemed "highly annoying but not painful" (e.g., Dunsmoor et al., 2016). The CSs were two angry faces (Ekman and Friesen, 1976) with a jittered stimulus duration of 6.5 ± 0.5 s, followed by a jittered waiting period with a fixation cross on a blank screen for 10.5 ± 1.0 s. Subjects were not instructed about the CS-US relationship at any point and had to learn this association for themselves.

Threat conditioning included a total of 18 CS$^+$ trials unpaired with the shock, 12 CS$^+$ trials paired with the shock (reinforcement rate of 40%), and 18 CS$^-$ trials. The first two trials (1 CS$^+$ and 1 CS$^-$) were considered habituation trials and were not included in the analysis. The third trial was always a CS$^+$ trial paired with shock. A 200 ms shock coterminated with the CS$^+$. Shocks were delivered to the right wrist using pregelled MRI-compatible electrodes connected to a stimulator (Grass Medical Instruments). Trials with the shock were removed from analysis in keeping with past protocols (e.g., Dunsmoor et al., 2016). Extinction included a total of 24 CS$^+$ and 24 CS$^-$ trials all unpaired with shock. During EXT, the shock was simply omitted. During NFE, the shock was omitted and each CS$^+$ trial instead coterminated with a low-volume 500 ms 440 Hz tone. The next day, shock electrodes were reattached, and subjects underwent a test of extinction–retention (12 CS$^+$ and 13 CS$^-$ trials) in the absence of the shock or the tone. The first trial during the 24 h test was always a CS$^-$ to account for initial orienting responses and was not included in analysis (compare Schiller et al., 2013; Kroes et al., 2017). The second and third trial were counterbalanced between subjects as either a CS$^+$ then a CS$^-$, or a CS$^-$ then a CS$^+$.

Psychophysiology. Autonomic arousal was measured throughout the scanning session by SCRs collected from pregelled MRI-compatible electrodes connected to the MP100 (BIOPAC Systems). Electrodes were attached to the hypothenar eminence of the left palm, and SCRs were calculated according to our previous criteria (Dunsmoor et al., 2015b). In brief, an SCR was considered related to stimulus presentation if the trough-to-peak deflection occurred within a time window that extended from 0.5 s following CS onset to the CS offset (jittered 6.5 ± 0.5 s), lasted between 0.5 and 5.0 s, and was >0.02 μS. If an SCR did not meet these criteria, then the trial was scored as 0. Responses were obtained using the MATLAB (The MathWorks) script Autonomate that extracted SCRs for each trial using these criteria (Green et al., 2014).

Subjects were not excluded from any analysis based on SCR results. That is, some prior studies have come under a bit of scrutiny for excluding subjects based on physiological performance (e.g., "poor learners" who do not reach a criterion of differential SCRs on CS$^+$ vs CS$^-$ trials during some phase of the experiment). To circumvent these issues, we did not exclude subjects based on SCR performance. This necessarily presents a trade-off whereby subjects who evinced weak SCRs (for whatever reason) are included along with subjects who evinced robust SCRs, but the alternative of selectively removing subjects presents its own drawbacks that potentially outweigh including all subjects into the analysis (for a thorough description of this issue, see Lonsdorf et al., 2017).

Computational modeling. In keeping with our prior computational model fitting approach (Raio et al., 2017), we fit and validated an associability model using individual subject SCRs. In the model, x_n indicates the CS on trial n (CS$^+$ or CS$^-$) and r_n is the US (1 for US, 0 for no US). Value (i.e., shock) predictions $V_n(x_n)$ were defined for each stimulus type (x_n) and trial. The prediction error $\delta_n = r_n - V_n(x_n)$ measures the difference between the expected and predicted shock on trial n. We replaced the constant learning rate from the Rescorla–Wagner model with a dynamic learning rate that gates the speed of learning based on a Pearce–Hall associability rule (Sutton, 1992; Le Pelley, 2004; Li et al., 2011; Raio et al., 2017). The resulting model for threat conditioning was as follows:

$$V_{n+1}(x_n) = V_n(x_n) + k\alpha_n(x_n)\delta_n$$

$$\alpha_{n+1}(x_n) = \eta_c|\delta_n| + (1 - \eta_c)\alpha_n(x_n)$$

where η_c indicates the weight assigned to the most recent absolute value of prediction error (indicating the accuracy of value prediction) in the conditioning phase and κ indicates a normalization factor. We reasoned that η might be modulated by replacing the aversive US with a novel neutral stimulus. We thus postulate that a new η might govern the updating of the rate of extinction learning (η_E) as follows:

$$\alpha_{n+1}(x_n) = \eta_E|\delta_n| + (1 - \eta_E)\alpha_n(x_n)$$

We tested the fit of this hybrid model by minimizing the difference between model-predicted associability and the skin conductance responses in both sessions. We optimized the free parameters of the model η_C, η_E *and k*) by maximizing the posterior probability of observing the measured sequence of SCRs following each CS. This maximization was achieved via the maximum *a posteriori* method that has been widely used in individual model fitting to avoid extreme parameter estimation.

The initial expected value V_0 and associability α_0 were set to 0.5, and prediction error weights (i.e., η_C, η_E) were constrained to the range (0, 1) with a β (1.2, 1.2) prior distribution slightly favoring values in the middle of the range; normalization factor κ was constrained to be positive values with a γ (1.2, 1) prior distribution. Log posterior probability was calculated by the summation of the observed data log-likelihoods and parameter log-priors for each subject. Further additional details on this model fitting procedure were reported previously (Li et al., 2011; Raio et al., 2017).

Imaging parameters and preprocessing. Whole-brain functional imaging was collected on a 3T Siemens Allegra head-only scanner at the Center for Brain Imaging at New York University. Preprocessing was conducted using SPM8 (Wellcome Trust Centre; www.fil.ion.ucl.ac.uk) implemented in MATLAB. Images were corrected for head motion using a 3 mm movement cutoff in any dimension. No subjects were removed for excessive head motion. Functional images were coregistered to each participant's high-resolution T1-weighted structural scan, spatially normalized into MNI space, voxel size resampled to 2 × 2 × 2 mm, and smoothed using an isotropic 8 mm^3 Gaussian FWHM kernel. The first four volumes of each functional run were discarded for T1 equilibrium.

Imaging analysis. Statistical analysis of preprocessed data was conducted using the GLM in SPM8. First-level (individual subject) analysis included covariates for the onset and offset of each CS modeled using an impulse function (0 s duration). Regressors for trial events were temporally convolved using a canonical hemodynamic response function. Covariates of no interest (i.e., nuisance regressors) included in the GLM were the CS$^+$ trials paired with shock and the shock itself, tones (for the NFE), 6 head-motion parameters, and estimates of signal intensity in white matter and CSF. A high-pass 128 s filter was applied to account for low-frequency drifts.

Second-level (group) analysis in SPM8 included a full factorial model with CS type (CS$^+$, CS$^-$) as the within-subjects factor and group (EXT, NFE) as the between-subjects factor. Statistical thresholds for whole-brain analyses were set at $p < 0.001$ with an extent threshold of 65 voxels, calculated using 10,000 Monte Carlo simulations in AlphaSim (Cox et al., 2017). Threshold for small-volume correction in the amygdala was set at FEW-corrected $p < 0.05$ using bilateral anatomical masks from the Pick-Atlas toolbox (Maldjian et al., 2003). Anatomical labels used in the tables for regions of activation were provided by Talairach Client (www.Talairach.org) (Lancaster et al., 2000) based on peak coordinates in MNI space, converted to Talairach space using GingerALE 2.3.6 (www.brainmap.org) (Laird et al., 2010). Any statistical test conducted outside SPM8 was from extracted mean β parameters and analyzed in SPSS 25 (IBM). *A priori* ROIs included those with a purported role in threat learning and expression (Fullana et al., 2016) and threat inhibition (Fullana et al., 2018b) as shown in extant fMRI research on Pavlovian conditioning and extinction; this included the insula, dACC, thalamus extending into midbrain, vmPFC, and lateral PFC. The amygdala was an *a priori* ROI as well, but it is not commonly observed in human fear conditioning fMRI (for meta-analysis, see Fullana et al., 2016), and did not emerge from the threat conditioning analysis, consistent with numerous failures to see amygdala activity in human fear conditioning fMRI (Fullana et al., 2018a).

Parametric modulation analysis of associability and prediction errors. To estimate regions tracking trial-by-trial changes in associability and prediction errors during the course of extinction learning, we used values derived from the Hybrid model (Li et al., 2011; Raio et al., 2017) as parametric regressors modulating CS onset (associability) and CS offset (prediction errors). For the first-level analysis, we concatenated all four runs from day 1 (2 runs of fear conditioning and 2 runs of extinction) and included session-specific constants (see also Boll et al., 2013) along with regressors for the shock, the 6 head-motion regressor, and estimates of signal intensity in white matter and CSF. To separately analyze associability and prediction error-related activity specific to extinction, we generated separate regressors for CS trials from threat conditioning and extinction. Regressors were convolved with the canonical hemodynamic response within the GLM at the first level and carried forward to the second level for whole-brain group level analysis using a threshold of $p <$ 0.001, cluster-corrected for multiple comparisons. Contrast estimates of associability and prediction error-related activity during extinction were derived using a one-sample *t* tests for each group separately, and a two-sample *t* test to find differences in activity between the groups.

Psychophysiological interaction (PPI). A PPI (Friston et al., 1997) was conducted using SPM8 to examine patterns of task-related functional connectivity during the 24 h retention test (day 2). The representative time course was extracted from source ROIs (described in Results) using a 4 mm sphere centered on the peak voxel within the ROI from the group-level analysis. The interaction term between the time series and the psychological context (PPI) was included in the GLM, using the CS$^+$ versus CS$^-$ as the trial regressor and including the same nuisance regressors as described above at the single-subject level. Whole-brain results were identified at $p < 0.001$ cluster-corrected for multiple comparisons, and a small-volume correction of FWE < 0.05 for the amygdala.

Results

Computational modeling dissociates standard and NFE

Conditioned SCRs (Fig. 1*B*) replicated prior findings that the NFE procedure accelerates extinction and diminishes conditioned SCRs at a 24 h extinction retention test (Dunsmoor et al., 2015c; Lucas et al., 2018). Given that the focus of this report is on the computational modeling and imaging results, we limit discussion of the SCR results. In short, both groups showed heightened SCRs to the CS$^+$ versus the CS$^-$ during threat conditioning (day 1), indicating successful acquisition of conditioned threat. Repeated-measures ANOVA using CS type (CS$^+$, CS$^-$) as a within-subjects factor and group (EXT, NFE) as a between-subjects factor showed a main effect of CS type ($F_{(1,44)} = 35.243$, $p < 0.001$, $\eta^2 = 0.445$), but not group ($p = 0.192$) and no CS type × group interaction ($p = 0.363$). Notably, while there was no difference in conditioning between groups, the EXT group on average showed nominally elevated SCRs to the CS$^+$, especially during late conditioning. This is likely due to considerable interindividual variability in SCRs, especially collected within a scanning environment. Notably, we included all subjects in the analysis, and did not exclude subjects on the basis of SCR performance.

The primary index of 24 h return of threat responses (day 2) was the mean SCR to the early (first four) CS$^+$ versus CS$^-$ trials. We focused on the mean of the early trials, rather than the entire testing phase, as this is the period of time when group differences in extinction retention tend to be most evident (Milad et al., 2009; Schiller et al., 2010; Kroes et al., 2016). After several unreinforced

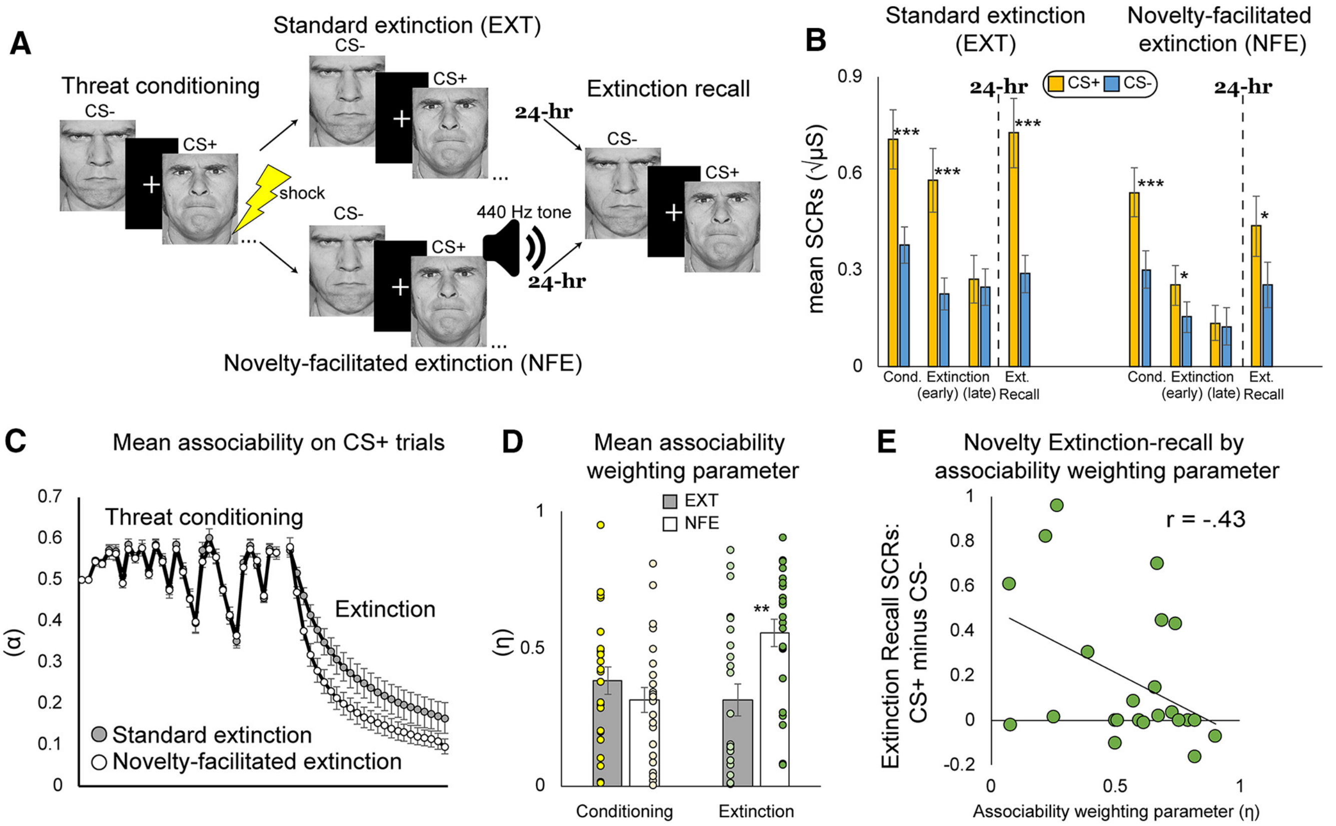

Figure 1. Experimental design and behavioral results. ***A***, Two groups underwent threat conditioning with a picture of a face (CS$^+$) paired with wrist shock on a partial CS-US pairing schedule, and a second picture (CS$^-$) not paired with shock. Conditioning was followed immediately by either EXT, in which the shock was omitted on CS$^+$ trials, or NFE, in which the shock was replaced by a tone at the end of each CS$^+$ trial. Subjects returned 24 h later, and the CSs were presented in the absence of any shocks or tones. ***B***, Conditioned SCRs replicate prior findings (Dunsmoor et al., 2015a), showing faster extinction and comparatively diminished SCRs 24 h later in the NFE group compared with the EXT group. ***C***, Best-fit associability trace for CS$^+$ trials illustrate accelerated updating during NFE compared with EXT. ***D***, The weighting parameter (η) that governs the rate of associability updating was elevated during NFE compared with EXT. ***E***, Individual differences in the associability weighting parameter during NFE was correlated with recovery of conditioned SCRs the next day, such that subjects who assigned more weight to the prediction error on NFE trials showed better retention of extinction 24 h later. Error bars indicate ± SEM. ***$p < 0.001$. **$p < 0.01$. *$p < 0.05$.

trials, subjects will begin to reextinguish, rendering it difficult to dissociate differences in extinction retention from variable rates of reextinction. The focus on early trials, and the first 4 trials in particular (e.g., Milad et al., 2009), is also in keeping with existing fMRI research using 24 h tests (Fullana et al., 2018b). Independent-samples *t* test on the SCR difference (CS$^+$ − CS$^-$) in both groups showed stronger differential SCRs in the EXT than NFE group ($t_{(44)} = 2.487$, $p = 0.017$, 95% CI [0.048, 0.460]). The EXT group showed substantially elevated SCRs on the CS$^+$ versus the CS$^-$ trials, as revealed by a one-sample *t* test on the CS$^+$ versus the CS$^-$ ($t_{(22)} = 5.632$, $p < 0.001$, 95% CI [0.276, 0.598]). Independent-samples *t* test on the CS$^+$ trials alone showed heightened mean SCRs in the EXT versus NFE group ($t_{(44)} = 2.015$, $p = 0.050$, 95% CI [−0.576, 0.0012]), but no difference in mean SCRs to the CS$^-$ ($p = 0.719$). While conditioned SCRs were diminished in the NFE group relative to the EXT group, conditioned arousal to the CS$^+$ at a 24 h test was not eliminated by the NFE procedure, and SCRs were elevated to CS$^+$ versus CS$^-$ ($t_{(22)} = 2.756$, $p = 0.012$, 95% CI [0.045, 0.321]). Together, these physiological results confirm that omitting the shock and replacing it with a nonaversive tone improved the rate of extinction learning led to comparatively less return of threat than just omitting the shock, replicating previous findings (Dunsmoor et al., 2015c).

To examine whether NFE affects trial-by-trial extinction learning differently than EXT, we fit an associability model using subject's trial-by-trial SCRs from day 1, in keeping with prior implementation of this model during human fear conditioning (Li et al., 2011; Raio et al., 2017). Repeated-measures ANOVA of the best-fit associability trace (α) on CS$^+$ trials (Fig. 1*C*) showed a main effect of group ($F_{(1,44)} = 5.062$, $p = 0.030$, $\eta^2 = 0.103$). Associability values were mostly undifferentiated between groups during threat conditioning for CS$^+$ trials but showed accelerated updating for the CS$^+$ during NFE compared with EXT. Repeated-measures ANOVA of the mean weighting factor (η) that governs the rate of trial-by-trial associability updating showed a phase (conditioning, extinction) × group interaction ($F_{(1,44)} = 10.708$, $p = 0.002$, $\eta^2 = 0.196$). Figure 1*D* shows that the mean weighting factor (η) was no different between groups during conditioning (independent-sample *t* test: $p = 0.31$), as expected given that the threat conditioning task was no different between groups. However, during extinction, the weighting factor η was significantly elevated in the NFE group versus the EXT group (independent-sample *t* test: $t_{(44)} = 3.150$, $p = 0.003$, 95% CI [0.088, 0.401]).

We also investigated whether individual differences in the associability weighting parameter η at the time of NFE (day 1) had any predictive value in determining the strength of extinction retention 24 h later (Fig. 1*E*). That is, we asked whether subjects who evince strong updating of associability during NFE on day 1 are those who show less return of threat responses on day 2. There was an inverse relationship between the associability weighting parameter at the time of NFE learning and differential ($CS^+ - CS^-$) SCRs at the time of test the next day ($r_{(21)} = -0.427$, $p = 0.04$, two-tailed). This suggests that those subjects for whom the surprising outcome on extinction trials most effectively increased associability were those who exhibited the strongest benefit, in terms of inhibiting future threat expression. There was no correlation between the associability weighting parameter derived from EXT and SCRs evinced 24 h later ($r_{(21)} = -0.177$, $p = 0.42$, two-tailed).

Imaging results

We first characterized whole-brain group-level fMRI analysis of threat conditioning to ensure equivalent threat conditioning-related activity in both groups. Across groups, threat conditioning ($CS^+ > CS^-$) was associated with activity in the bilateral insula, striatum, thalamus extending into the midbrain, and dACC (Table 1). The inverse contrast ($CS^+ < CS^-$) revealed activity in the vmPFC, posterior cingulate cortex, and left angular gyrus. These regions are consistently implicated in fMRI studies of human conditioning (Sehlmeyer et al., 2009; Mechias et al., 2010; Fullana et al., 2016). Importantly, no regions emerged as showing a main effect of group, or a group × CS type interaction, confirming that activity related to initial acquisition was equivalent between groups, as expected.

Whole-brain analysis of extinction

Repeated-measures 2 × 2 ANOVA of whole-brain activity during extinction revealed a main effect of CS type in a number of areas that were also active during threat conditioning, including the dACC, bilateral insula, and striatum (for full report, see Table 2). This overlap in activity within threat learning and expression networks across conditioning and extinction has been reported in a meta-analysis of human fear extinction fMRI (Fullana et al., 2018b). The group × CS type interaction revealed group differences in the vmPFC, dACC, insula, and superior frontal gyrus (Fig. 2*A*; Table 3). To characterize the interaction, we extracted parameter estimates for the CS^+ and CS^- contrast from *a priori* ROIs (Fig. 2*A*). This confirmed that the interaction in ROIs associated with threat inhibition (the vmPFC and dorsolateral PFC; Fig. 2*A*,*B*) threat appraisal and expression (i.e., the insula and dACC; Fig. 2*C*,*D*) were driven by heightened differential activity between the CS^+ and CS^- during EXT compared with NFE. Notably, in the vmPFC and lateral PFC, the pattern in activity between the CS^+ and CS^- reversed from conditioning to extinction, such that deactivation to the CS^+ was diminished. Thus, despite equivalent acquisition-related activity, NFE produced a dramatic and highly distinguishable pattern of BOLD fMRI in regions involved in threat inhibition and threat appraisal and expression.

Associability-related activity in the vmPFC during NFE

We compared associability-related brain activity during extinction by using parametric modulation analyses of trial-by-trial associability values for each participant. To detect differences in associability signals at the whole-brain level, we conducted a two-sample *t* test on the associability regressor from the parametric modulation analysis. This revealed stronger associability-modulated activity in the vmPFC during NFE compared with EXT (Fig. 2*E*). This was the only region identified that survived whole-brain correction for multiple comparisons. A two-sample *t* test on the prediction-error regressor for extinction did not reveal any differences between groups.

Whole-brain analysis of retention test

A whole-brain 2 × 2 ANOVA of the 24 h test phase showed a group × CS type interaction in the vmPFC, a region of the vmPFC consistent with the subgenual ACC (sgACC, Brodmann area 25), and bilateral posterior parietal cortex (Table 4). Further analysis of the vmPFC confirmed that differences between CS^+ and CS^- in the EXT were driven by deactivations to the CS^+ relative to the CS^-, whereas the NFE showed a relative increase in activity to the CS^+ that was near the baseline level of the learned safety signal, the CS^- (Fig. 3*A*). The main effect of CS type also revealed activity in a number of regions that overlapped with areas identified in threat conditioning (e.g., bilateral insula, thalamus, striatum, ACC, posterior cingulate cortex, vmPFC; Table 5). Interestingly, this analysis also revealed activity in the left amygdala (MNI −18, 0, −12; 7 voxels, $F = 17.62$, $P_{\text{FWE-corrected}} = 0.01$). Given the general interest in the role of the amygdala in extinction processes (Ehrlich et al., 2009; Pape and Paré, 2010), we examined activity in this amygdala ROI during the 24 h retention test and found that both groups exhibited enhanced CS^+ versus CS^- responses in the left amygdala (Fig. 3*B*). This amygdala ROI was used as a seed region for a subsequent functional connectivity analysis reported below.

Amygdala-vmPFC connectivity characterizes recall of NFE versus EXT

The vmPFC is often associated with successful extinction recall in humans (Phelps et al., 2004; Milad et al., 2007) and corresponds to an area of the mPFC consistent with the infralimbic cortex in rodents that is generally implicated in threat extinction processes (Quirk and Mueller, 2008). Connectivity patterns between the mPFC and amygdala have also been characterized in research on inhibitory networks controlling emotional and defensive responding in humans and rodents (Duvarci and Paré, 2014; Frank et al., 2014; Tovote et al., 2015; but see Bukalo et al., 2015). To further probe the role of this region during recall of standard versus NFE, we used this ROI from the extinction–retention test as a seed region in a whole-brain functional connectivity analysis (PPI). At the whole-brain level, using a somewhat liberal threshold of $p < 0.001$ uncorrected and a voxel extent of 5 voxels, the only two regions to emerge as exhibiting stronger task-based ($CS^+ > CS^-$) functional correlations with the vmPFC in the NFE versus EXT group during extinction recall were in right (MNI −20, −2, −22; 15 voxels, $t = 3.72$, $P_{\text{uncorrected}} < 0.001$) and left (MNI 26, −4, −24; 8 voxels, $t = 3.73$, $P_{\text{uncorrected}} < 0.001$) amygdala (Fig. 3*C*).

A secondary and complementary analysis used the left amygdala ROI identified from the main effect of CS type across both groups during extinction recall as the PPI seed region. At the whole-brain level, the only region to emerge as showing stronger positive correlations with the amygdala during the 24 h test in the NFE group versus the EXT group was the vmPFC (Fig. 3*D*), perhaps unsurprisingly mirroring the results using the vmPFC as the PPI seed region. At a more lenient exploratory threshold of $p < 0.005$, we found that the EXT group exhibited stronger amygdala connectivity with the dACC compared with the NFE group (two-samples *t* test).

Discussion

Behavioral strategies can enhance extinction to prevent relapse of extinguished behaviors (Craske et al., 2008, 2014; Laborda et al., 2011; Dunsmoor et al., 2015a). However, despite increasing knowledge on the neuroscience of fear extinction, the neural cor-

Table 1. Whole-brain ANOVA of CS type (CS^+, CS^-) and group (EXT, NFE), identified at $p < 0.001$ (cluster-corrected $p < 0.05$) during threat (fear) conditioning across all participants ($n = 46$)

	MNI coordinate					
Region	*x*	*y*	*z*	Size (voxels)	Peak *T*	Peak *Z*
$CS^+ > CS^-$						
Postcentral gyrus	−60	−22	22	1750	10.09128	
Insula	−50	−30	22		9.339011	7.757533
Insula	−42	−24	16		4.907147	4.600763
Precentral gyrus	−46	2	6	4659	9.422381	7.804049
Insula	−36	10	6		9.011673	7.563284
Insula	−30	26	4		7.629888	6.666249
Insula	48	10	−2	2719	8.610116	7.313175
Precentral gyrus	56	10	2		7.573494	6.627522
Insula	32	20	−12		7.055649	6.264065
Cingulate gyrus	−6	12	38	2686	8.599445	7.306483
Cingulate gyrus	0	8	42		8.172343	7.030367
Cingulate gyrus	−4	2	48		7.691855	6.708615
Precentral gyrus	46	0	44	408	7.348836	6.471585
Middle frontal gyrus	52	6	42		7.200085	6.366869
Precentral gyrus	−38	−8	50	271	6.867754	6.128654
Precentral gyrus	−50	−2	44		4.731482	4.453624
Insula	58	−28	20	1397	6.693943	6.001698
Insula	60	−36	22		6.502241	5.859773
Superior temporal gyrus	50	−40	14		4.981651	4.662666
Superior frontal gyrus	−36	46	26	196	5.533299	5.111623
Precuneus	−18	−52	58	147	5.239964	4.874961
Inferior parietal lobule	−34	−42	56		4.85981	4.561275
Precuneus	−24	−44	56		3.949349	3.778787
Superior temporal gyrus	50	−24	−10	175	4.752214	4.471075
Superior temporal gyrus	56	−30	0		3.826558	3.670006
Superior frontal gyrus	26	50	20	222	4.2548	4.046103
Superior frontal gyrus	32	54	24		4.01749	3.83883
Medial frontal gyrus	26	42	22		3.998741	3.822332
Caudate body	12	4	4	189	4.21751	4.013723
Lateral globus pallidus	20	2	−2		4.100237	3.911429
$CS^+ < CS^-$						
Posterior cingulate	−2	−56	20	1861	6.757084	6.048007
Precuneus	−4	−60	30		6.133412	5.581067
Cingulate gyrus	−6	−40	34		5.337781	4.954403
Precuneus	−32	−72	42	1495	6.399227	5.782679
Middle temporal gyrus	−46	−64	28		6.379358	5.767743
Middle temporal gyrus	−38	−62	36		4.879001	4.577298
Superior frontal gyrus	−20	62	12	3401	6.149881	5.593671
Medial frontal gyrus	−4	50	−18		5.936932	5.42955
Medial frontal gyrus	−6	56	−6		5.73744	5.273538
Inferior frontal gyrus	−44	52	−4	136	5.631363	5.189689
Middle temporal gyrus	40	−68	34	995	5.573978	5.144071
Middle temporal gyrus	46	−60	24		5.226706	4.864153
Middle temporal gyrus	40	−54	28		5.017684	4.692497
Middle temporal gyrus	−62	−40	−8	162	5.149294	4.800857
Middle frontal gyrus	48	36	12	112	5.101435	4.761562
Postcentral gyrus	32	−22	42	286	4.911562	4.604439
Precentral gyrus	30	−24	52		4.702389	4.429096
Postcentral gyrus	56	−14	40		4.105146	3.915724
Middle temporal gyrus	−58	−14	−20	68	4.227094	4.022052
Subgyral	22	34	−16	75	4.00586	3.828598
Inferior frontal gyrus	32	36	−12		3.545974	3.418684
Cuneus	14	−90	4	63	24.07	4.45
Main effect of group						
No regions	—	—	—	—	—	—
Group × CS type interaction						
No regions	—	—	—	—	—	—

relates underlying behaviorally enhanced extinction strategies are unclear. Here we found that enhanced extinction that involves replacing threats with novel outcomes (Dunsmoor et al., 2015c) diminished activity in threat appraisal and expression regions, including the insula, thalamus, and dACC. A computational learning model that emphasizes surprise-oriented attention and dynamically governs how easily the CS can form an association with its outcome characterized vmPFC activity during NFE, and individual differences in associability-modulated extinction correlated with diminished conditioned arousal the next day. Finally, vmPFC activity was functionally correlated with the amygdala during 24 h later, in line with rodent neurobiological

Table 2. Whole-brain ANOVA of CS type (CS $^+$, CS $^-$) and group (EXT, NFE), identified at $p < 0.001$ (cluster-corrected $p < 0.05$) during threat (fear) extinction across all participants ($n = 46$)

	MNI coordinate					
Region	*x*	*y*	*z*	Size (voxels)	Peak *T*	Peak *Z*
CS $^+$ > CS $^-$						
Inferior frontal gyrus	44	24	2	2401	8.909137	7.500645
Precentral gyrus	46	12	2		8.30936	7.11996
Insula	40	16	6		7.576684	6.629717
Insula	−34	20	8	4273	8.805809	7.436392
Superior temporal gyrus	−54	4	4		8.165496	7.025867
Precentral gyrus	−46	4	6		7.010934	6.232012
Inferior parietal lobule	−64	−28	24	933	7.321795	6.452637
Supramarginal gyrus	−56	−44	34		6.225281	5.651185
Cingulate gyrus	−4	0	46	3236	6.910562	6.159672
Cingulate gyrus	6	18	38		6.81266	6.088589
Cingulate gyrus	4	22	30		6.795955	6.076408
Precentral gyrus	46	0	42	317	6.431106	5.806599
Middle frontal gyrus	−32	46	32	382	6.270257	5.685344
Middle frontal gyrus	−36	50	24		5.612369	5.174609
Insula	54	−38	26	579	6.087015	5.545479
Superior temporal gyrus	64	−38	24		5.192222	4.835997
Supramarginal gyrus	58	−48	30		5.015621	4.690791
Precentral gyrus	−36	−6	50	165	5.837435	5.352012
Precentral gyrus	−24	−10	52		4.219384	4.015352
Superior frontal gyrus	26	46	24	191	5.001156	4.678823
Superior frontal gyrus	26	58	26		4.082089	3.895536
Precuneus	−12	−70	42	65	4.514843	4.269898
Parahippocampal gyrus	50	−30	−12	60	4.121172	3.929741
Middle temporal gyrus	52	−34	−4		3.511212	3.387287
CS $^+$ < CS $^-$						
Cuneus	22	−82	34	6488	7.017042	6.236396
Precuneus	−30	−74	38		6.730351	6.028427
Middle temporal gyrus	44	−64	26		6.382114	5.769817
Middle frontal gyrus	50	40	14	100	4.924735	4.615403
Inferior frontal gyrus	50	42	6		3.964584	3.792231
Middle frontal gyrus	46	48	12		3.824497	3.668174
Parahippocampal gyrus	−28	−44	−12	376	4.634379	4.371582
Culmen	−18	−64	−10		4.304164	4.088857
Parahippocampal gyrus	−24	−50	−8		4.002995	3.826077
Middle temporal gyrus	−58	−44	−10	236	4.605868	4.347397
Middle temporal gyrus	−64	−38	−2		4.328274	4.109692
Middle temporal gyrus	−54	−36	−4		3.642833	3.505867
Middle temporal gyrus	−62	−12	−14	132	4.576139	4.322134
Superior frontal gyrus	−22	22	44	101	4.568189	4.31537
Middle frontal gyrus	−26	14	46		4.155802	3.959984
Subgyral	−18	30	40		4.092723	3.90485
Medial frontal gyrus	−10	50	32	244	4.566406	4.313853
Superior frontal gyrus	−10	64	14		4.230076	4.024642
Superior frontal gyrus	−14	46	38		3.362291	3.25215
Paracentral lobule	6	−30	56	60	4.562235	4.310302
Paracentral lobule	−2	−30	58		3.888207	3.724714
Middle frontal gyrus	−44	24	24	224	4.52998	4.282816
Medial frontal gyrus	−10	44	−14	160	4.447126	4.21196
Medial frontal gyrus	2	54	−10		3.764548	3.61479
Precentral gyrus	34	−24	52	89	4.396607	4.16858
Precentral gyrus	34	−16	46		4.076615	3.890738
Postcentral gyrus	36	−32	54		3.536506	3.410139
Caudate body	28	−16	26	70	4.303317	4.088123
Superior temporal gyrus	60	−4	−16	80	3.974018	3.800551
Superior temporal gyrus	66	−10	2		3.926633	3.758719
Superior temporal gyrus	66	−12	−8		3.889014	3.725429
Postcentral gyrus	54	−22	38	105	3.86692	3.705845
Postcentral gyrus	58	−12	38		3.691826	3.549797
Postcentral gyrus	54	−22	48		3.416973	3.301888

models emphasizing prefrontal-amygdala pathways regulating the balance of threat and safety.

The whole-brain imaging results showed that enhancing extinction reduced BOLD activity in a number of regions associated with threat appraisal and expression, including the insula, dACC, and thalamus. Activity to the CS $^+$ is traditionally maintained in these regions during EXT learning in human fMRI (Fullana et al., 2018b), perhaps because extinction tends to be slow and the mere omission of shock is not sufficient to reduce neural activity related to threat anticipation. It may be especially difficult to reduce

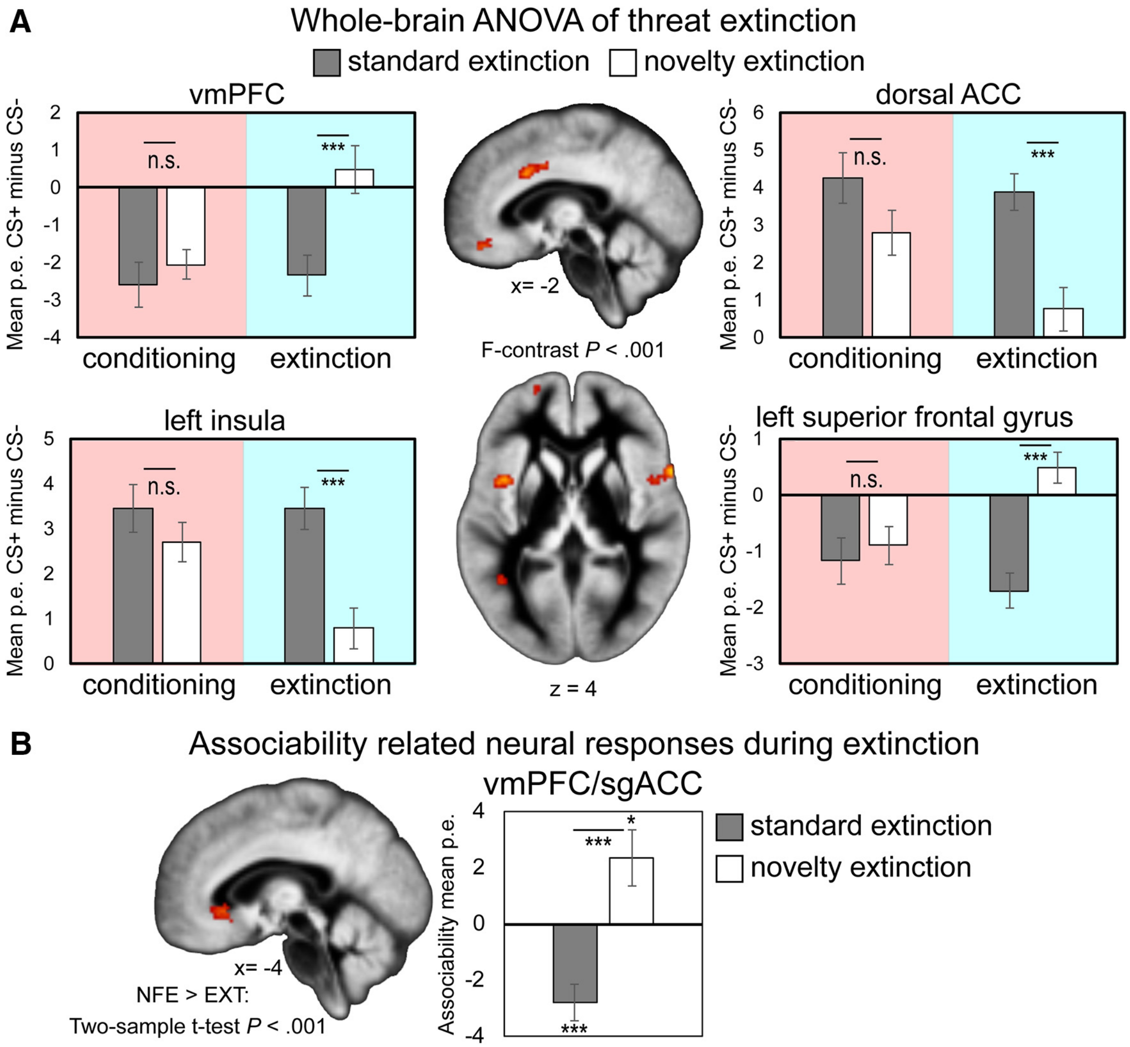

Figure 2. Whole-brain ANOVA and associability-modulated vmPFC activity during extinction. The group × CS type interaction of extinction revealed activations in vmPFC, superior frontal gyrus, dACC, and insula. ***A***, Parameter estimates extracted from these regions characterized the interaction as deactivations to CS$^+$ versus CS$^-$ in vmPFC and superior frontal gyrus during both conditioning and extinction in the EXT group but a switch in CS$^+$ versus CS$^-$ activity during extinction in the NFE group. dACC and left insula exhibited heightened CS$^+$ versus CS$^-$ differential activity during condition and extinction in the EXT group but diminished CS$^+$ versus CS$^-$ activity during extinction in the NFE group. ***B***, NFE exhibited stronger associability-modulated engagement of the vmPFC than EXT. This region of the vmPFC corresponds to an area of the sgACC considered homologous to the rodent infralimbic cortex. p.e., Parameter estimates, arbitrary units. Error bars indicate ± SEM. ***$p < 0.001$ (one-sample t test). *$p < 0.05$ (one-sample t test).

threat anticipation through passive omission of the US following partial CS-US pairing (Grady et al., 2016). However, replacing threat outcomes on extinction trials appears to effectively attenuate activity in regions that are otherwise involved in maintaining and expressing threat expectations. One psychological effect of replacing the shock with a perceptible outcome (rather than just passively omitting it) might involve reducing threat uncertainty (Dunsmoor et al., 2015c; Morriss et al., 2016; Lucas et al., 2018).

Computational modeling showed that associability modulated vmPFC activity during NFE. Given that extinction is considered new associative learning, it is intriguing that a region involved in extinction learning (the vmPFC) was engaged by dynamic changes in associability. This suggests that novelty modulates the effectiveness of new associative learning in the vmPFC. In this framework, the vmPFC might provide a "teaching signal" that serves to instruct the formation of extinction memories (Bukalo et al., 2015) while simultaneously downregulating regions involved in maintaining threat appraisal and emotional expression, such as the insula, thalamus, brainstem, and dACC. This explanation fits with studies in rodents showing that electrical (Milad et al., 2004) or optogenetic (Do-Monte et al., 2015) stimulation of the infralimbic cortex facilitates extinction learning and strengthens extinction memories. This also fits with an investigation in rodents showing that replacing a shock US with an appetitive US increases activity in the infralimbic cortex and decreases long-term defensive responses (Correia et al., 2016). It is noteworthy, therefore, that the region of the vmPFC identified as tracing associability better in NFE than EXT was the sgACC (Brodmann area 25). This area is commonly thought of as homologous to infralimbic cortex (Ongür et al., 2003), which is crucial for extinction learning (Milad and Quirk, 2012).

A consequence of enhanced extinction learning might be the formation of a more durable extinction memory, which then provides stronger retrieval competition against the original threat association at test (Miller and Laborda, 2011; Laborda and Miller, 2012). This idea of strengthening retrieval competition is further supported by patterns of amygdala-vmPFC connectivity 24 h later. Specifically, activity between the vmPFC and amygdala was positively correlated on CS$^+$ versus CS$^-$ trials 24 h after NFE, whereas activity between the dACC and amygdala was pos-

Table 3. Whole-brain ANOVA of main effects of group (EXT, NFE) and CS type (CS$^+$, CS$^-$) × group interactions, during threat (fear) extinction, identified at $p < 0.001$ (cluster-corrected $p < 0.05$)

	MNI coordinate					
Region	*x*	*y*	*z*	Size (voxels)	Peak *F*	Peak *Z*
Main effect of group						
Cingulate gyrus	2	−50	30	438	26.22985	4.636714
Precuneus	−2	−54	42		16.58164	3.715015
Precuneus	−6	−60	28		16.2062	3.67262
Subgyral	−42	−28	4	364	24.21592	4.466096
Superior temporal gyrus	−52	−24	2		20.21357	4.09596
Superior temporal gyrus	−62	−20	4		19.41432	4.016267
Superior temporal gyrus	54	−12	2	405	19.8485	4.059824
Superior temporal gyrus	66	−10	4		19.7693	4.051926
Superior temporal gyrus	60	−30	2		19.30035	4.004727
Superior temporal gyrus	52	−56	16	63	19.50166	4.025081
CS × group interaction: [NFE (CS$^+$ > CS$^-$) > EXT (CS$^+$ > CS$^-$)]						
Middle frontal gyrus	−22	34	36	288	4.892444	4.588511
Subgyral	−24	14	40		3.885337	3.722172
Subgyral	−20	6	48		3.605171	3.472021
Superior temporal gyrus	50	−56	18	201	4.550718	4.300495
Middle temporal gyrus	40	−62	28		3.616361	3.482084
Subgyral	36	−54	28		3.580528	3.449838
Anterior cingulate	2	44	−16	155	4.223762	4.019156
Anterior cingulate	−6	42	−16		3.681697	3.540724
Anterior cingulate	2	36	−12		3.591407	3.459634
Middle frontal gyrus	26	32	36	60	4.087843	3.900577
Precentral gyrus	−46	26	32	63	3.589665	3.458066
Middle frontal gyrus	−48	32	26		3.38397	3.271886
CS × group interaction: [EXT (CS$^+$ > CS$^-$) > NFE (CS$^+$ > CS$^-$)]						
Inferior parietal lobule	50	−28	26	486	5.579247	5.148266
Inferior parietal lobule	66	−32	26		4.382391	4.156349
Inferior parietal lobule	54	−36	26		4.348459	4.127112
Precentral gyrus	60	12	4	110	4.518246	4.272803
Insula	48	8	0		3.969444	3.796518
Insula	−40	6	4	107	4.305992	4.090437
Cingulate gyrus	−2	14	32	247	4.220178	4.016042
Cingulate gyrus	−2	2	38		3.799844	3.646242
Cingulate gyrus	2	22	28		3.714503	3.570091
Inferior parietal lobule	−62	−36	30	252	4.214053	4.010718
Superior temporal gyrus	−64	−34	18		3.897533	3.732974
Inferior parietal lobule	−56	−28	24		3.604238	3.471181

Table 4. Whole-brain ANOVA of main effects of group (EXT, NFE) and CS type (CS$^+$, CS$^-$) × group interactions, during 24 h extinction recall, identified at $p < 0.001$ (cluster-corrected $p < 0.05$)

	MNI coordinate					
Region	*x*	*y*	*z*	Size (voxels)	Peak *F*	Peak *Z*
Main effect of group						
No regions	—	—	—	—	—	—
CS × group interaction: [NFE (CS$^+$ > CS$^-$) > EXT (CS$^+$ > CS$^-$)]						
Anterior cingulate	−6	26	−4	77	4.722658	4.446189
Anterior cingulate	2	24	−12		3.44957	3.331474
Middle temporal gyrus	66	−30	−6	61	4.166023	3.968899
Superior temporal gyrus	58	−58	20	144	3.990917	3.815443
Superior temporal gyrus	56	−46	12		3.494784	3.37243
Poster cingulate	−12	−54	26	63	3.917257	3.750429
Middle temporal gyrus	−48	−64	28	188	3.915658	3.749014
Middle temporal gyrus	−42	−66	34		3.594918	3.462794
Middle temporal gyrus	−38	−56	26		3.563415	3.434415
Anterior cingulate	−12	48	−16	176	3.851778	3.692409
Anterior cingulate	−6	46	−8		3.831742	3.674614
Medial frontal gyrus	4	56	8	64	3.838834	3.680915
Medial frontal gyrus	−4	56	12		3.554726	3.42658
CS × group interaction: [EXT (CS$^+$ > CS$^-$) > NFE (CS$^+$ > CS$^-$)]						
Frontal lobe/subgyral	34	12	20	63	4.386405	4.159803
Inferior parietal lobule	−42	−40	54	216	4.377207	4.151887
Supramarginal gyrus	−54	−40	42		4.139938	3.946138
Inferior parietal lobule	−52	−34	52		3.315179	3.20919

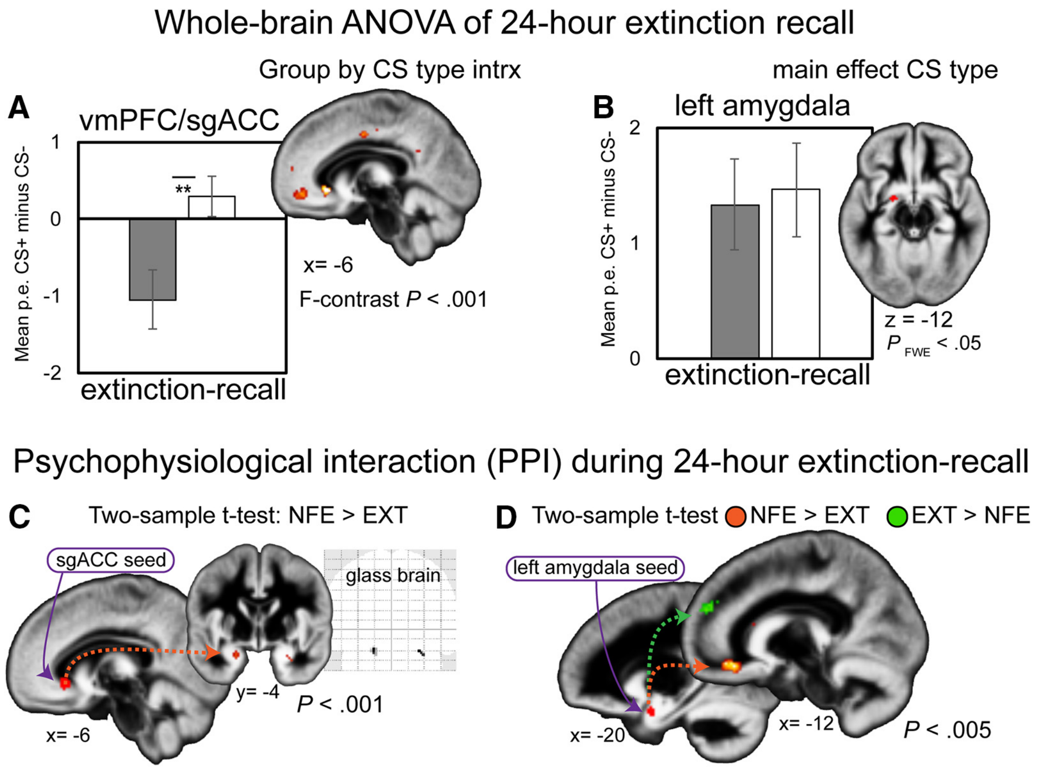

Figure 3. Whole-brain ANOVA and functional connectivity at 24 h test. ***A***, The group × CS type interaction revealed activity in the vmPFC, including a region in the sgACC. Parameter estimates extracted from the vmPFC/sgACC characterized the interaction as strong CS$^+$ deactivations 24 h after EXT compared with NFE. ***B***, The main effect of CS type revealed activation in left amygdala. ***C***, Task-based functional connectivity analysis using the vmPFC/sgACC as a seed region showed stronger correlations between the vmPFC and amygdala 24 h after NFE compared with EXT. ***D***, A complementary exploratory analysis using the left amygdala as a seed region showed dissociable patterns of connectivity between groups, with the NFE group exhibiting stronger connectivity with the vmPFC (mirroring the prior analysis) and the EXT group exhibiting stronger connectivity with the dACC 24 h later. p.e., Parameter estimates, arbitrary units. Error bars indicate ± SEM. $^{**}p < 0.01$.

itively correlated 24 h after EXT (albeit at a more liberal statistical threshold). These connectivity results fit with rodent neurobiological research showing that distinct pathways between the amygdala and the ventral (infralimbic) and dorsal (prelimbic) regions of the mPFC mediate the balance between inhibition and expression of conditioned threat (Sierra-Mercado et al., 2011; Senn et al., 2014; Krabbe et al., 2018). A positive correlation between the vmPFC and amygdala might be related to prefrontal inhibition via excitatory inputs to the basolateral amygdala (Bloodgood et al., 2018) that in turn activate GABAergic intercalated cell masses that project onto and inhibit the central nucleus of the amygdala (Amano et al., 2010; Strobel et al., 2015), although it has to be noted that fMRI methods are unable to substantiate that speculation. Overall, enhancing extinction might engage the vmPFC to strengthen new associative learning during extinction, which may then lead to stronger interference with the threat memory at test (Bouton, 1993).

Further research testing the NFE protocol, or methodological offshoots, is warranted. For instance, the present study and prior behavioral study (Dunsmoor et al., 2015c) used an immediate extinction design in humans, although the NFE protocol in rats in Dunsmoor et al. (2015c) did incorporate a 3 d delayed extinction and 24 h extinction retention test. Immediate extinction may be a weaker form of extinction that produces less extinction retention than delaying extinction by at least a day (Maren, 2014). It is perhaps unsurprising then that the EXT group exhibited enhanced activity to the CS$^+$ versus CS$^-$ throughout extinction, and showed weak extinction retention. At the same time, it is noteworthy that the NFE protocol reduced activity in regions associated with threat appraisal and improved extinction retention, despite the immediate extinction design. It will be of interest to test effects of an immediate versus delayed NFE protocol, as its possible NFE is even more effective at a delay, in keeping with findings of EXT standard extinction (Maren and Chang, 2006; Schiller et al., 2008; Huff et al., 2009). This study also used threat-relevant CSs (angry faces) that are more difficult to extinguish than neutral CSs (Ohman and Mineka, 2001). This again seems to support the idea that NFE is a more optimal technique that overcomes deficits in extinction to threat-relevant CSs, but it might be of interest to compare NFE to EXT of neutral CSs in humans. It also remains important to probe the similarity and differences between NFE, which incorporates neutral outcomes, and counterconditioning, which incorporates rewarding outcomes. We have speculated previously (Dunsmoor et al., 2015c) that NFE might be more effective than counterconditioning by effectively neutralizing the emotional significance of the CS, but there is also evidence that rewarded extinction is effective at diminishing future threat through an amygdala-ventral striatum circuit (Correia et al., 2016). Where these procedures overlap and differ is of theoretical interest. It is possible that the neutral tone used here to replace the shock acquires rewarding properties by being associated with omission of shock.

These results have implications for treatments of pathological anxiety based on associative learning theory. Exposure therapy is informed in large measure by basic principles of experimental extinction (Foa et al., 1989), and advances in the neuroscience of extinction offer potential insight into treatment for fear and anxiety disorders. Standard models of extinction rely on the repeated

Table 5. Whole-brain ANOVA of the main effect of CS type (CS$^+$, CS$^-$) identified at $p < 0.001$ (cluster-corrected $p < 0.05$) during 24 h extinction recall across all participants ($n = 46$)

	MNI coordinate					
Region	*x*	*y*	*z*	Size (voxels)	Peak *F*	Peak *Z*
Main effect of CS type (small-volume correction of bilateral amygdala at FWE < 0.05)						
Amygdala	−18	0	−12	7	17.62	3.83
CS$^+$ > CS$^-$						
Cingulate gyrus	0	18	28	2545	8.758961	7.407186
Cingulate gyrus	4	8	44		8.486077	7.234068
Medial frontal gyrus	−4	−4	50		5.927033	5.42186
Insula	40	22	0	2774	8.357562	7.151239
Claustrum	30	24	−2		6.814871	6.0902
Precentral gyrus	56	12	2		6.480037	5.843205
Claustrum	−30	22	4	2623	7.869288	6.828816
Precentral gyrus	−48	2	4		7.586207	6.636267
Claustrum	−40	12	−2		6.703367	6.008623
Insula	−58	−32	24	1286	6.76322	6.052496
Inferior parietal lobule	−56	−40	30		6.276483	5.690063
Transverse temporal gyrus	−46	−22	12		4.445009	4.210145
Caudate body	10	8	2	407	6.500726	5.858643
Thalamus	10	−2	−2		4.726253	4.449218
Thalamus	10	−16	0		4.627984	4.366161
Precentral gyrus	46	2	42	391	6.139207	5.585504
Middle frontal gyrus	36	−4	46		5.190931	4.834943
Precuneus	−26	−40	54	120	6.018391	5.492625
Precentral gyrus	−36	−6	50	258	5.673821	5.223324
Precentral gyrus	−50	0	46		5.195867	4.838976
Precentral gyrus	−52	2	38		4.533932	4.286187
Lateral globus pallidus	−18	0	−8	225	5.341197	4.957168
Caudate body	−10	6	2		4.167788	3.970437
Inferior parietal lobule	64	−26	22	810	4.940672	4.628655
Superior temporal gyrus	66	−36	24		4.845015	4.548909
Inferior parietal lobule	56	−28	22		4.741972	4.462457
Middle frontal gyrus	−34	38	28	222	4.879611	4.577807
Superior frontal gyrus	28	50	28	255	4.513252	4.26854
Medial frontal gyrus	24	44	16		4.055169	3.87193
Middle frontal gyrus	40	46	28		3.57616	3.445902
Thalamus	−8	−20	2	243	4.232623	4.026855
Midbrain	−4	−26	−4		4.146053	3.951476
Midbrain	−4	−28	−12		4.017396	3.838747
CS$^+$ < CS$^-$						
Posterior cingulate	2	−60	18	2816	8.099367	6.982266
Culmen	12	−46	0		5.253232	4.885767
Precuneus	0	−46	34		5.218146	4.857171
Precuneus	−32	−74	42	1297	6.620217	5.947352
Middle temporal gyrus	−44	−74	34		6.443248	5.815695
Cuneus	−26	−84	32		4.906551	4.600266
Middle temporal gyrus	50	−70	28	1332	6.315139	5.71932
Cuneus	30	−80	38		5.814116	5.333761
Precuneus	34	−72	42		5.373721	4.983461
Anterior cingulate	−4	42	−12	1895	5.760348	5.291564
Superior frontal gyrus	−10	60	18		5.758396	5.290029
Superior frontal gyrus	−12	60	30		4.7949	4.506935
Middle temporal gyrus	−56	−10	−16	160	5.004987	4.681994
Superior temporal gyrus	−50	0	−14		3.34302	3.23459
Hippocampus	28	−18	−22	76	4.567359	4.314664
Parahippocampal gyrus	26	−38	−14	74	4.409862	4.179975
Declive	−18	−70	−12	108	3.877688	3.715393
Declive of vermis	0	−76	−6		3.724742	3.579246
Lingual gyrus	−12	−74	−6		3.283707	3.180436

absence of the outcome to eventually produce a new CS association, at which point cue associability diminishes as the absence of the outcome is fully predicted (Pearce and Hall, 1980). Yet learning by omission seems to be a rather passive strategy to generate a competitive long-term extinction memory, as evidenced by a century of research on postextinction recovery of conditioned behaviors (Pavlov, 1927; Bouton, 2002; Rescorla, 2004). The imbalance between conditioning and extinction is especially evident in threat learning, where there is an evolutionarily conserved bias toward anxiety conservation (Solomon and Wynne, 1954; Bateson et al., 2011) to ensure that threat associations are maintained "just in case." Likewise, in many real-world situations, the mere absence of an expected threat is not sufficient to reduce threat uncertainty, exemplified in specific phobias (e.g., fear of flying: "just because the plane didn't crash this time doesn't mean it won't crash next time."). Extinction strategies that rely on passive

US omission might simply be insufficient in some cases (e.g., following partial reinforcement) (Li et al., 2016) or in some populations (e.g., stress disorders) (Milad et al., 2009) to stimulate vmPFC involvement. An inability to properly learn from experiences that disconfirm negative expectations could also explain why humans have more difficulty retrieving explicit memories of safety after extinction (Dunsmoor et al., 2018). Amplifying the effects of surprise by replacing expected threat with novel outcomes may be a potential technique to strengthen new safety memories to outcompete reactivation of unwanted negative memories.

References

Amano T, Unal CT, Paré D (2010) Synaptic correlates of fear extinction in the amygdala. Nat Neurosci 13:489–494.

Ball TM, Knapp SE, Paulus MP, Stein MB (2017) Brain activation during fear extinction predicts exposure success. Depress Anxiety 34:257–266.

Bateson M, Brilot B, Nettle D (2011) Anxiety: an evolutionary approach. Can J Psychiatry 56:707–715.

Bloodgood DW, Sugam JA, Holmes A, Kash TL (2018) Fear extinction requires infralimbic cortex projections to the basolateral amygdala. Transl Psychiatry 8:60.

Boll S, Gamer M, Gluth S, Finsterbusch J, Büchel C (2013) Separate amygdala subregions signal surprise and predictiveness during associative fear learning in humans. Eur J Neurosci 37:758–767.

Bouton ME (1993) Context, time, and memory retrieval in the interference paradigms of pavlovian learning. Psychol Bull 114:80–99.

Bouton ME (2002) Context, ambiguity, and unlearning: sources of relapse after behavioral extinction. Biol Psychiatry 52:976–986.

Bukalo O, Pinard CR, Silverstein S, Brehm C, Hartley ND, Whittle N, Colacicco G, Busch E, Patel S, Singewald N, Holmes A (2015) Prefrontal inputs to the amygdala instruct fear extinction memory formation. Sci Adv 1:e1500251.

Correia SS, McGrath AG, Lee A, Graybiel AM, Goosens KA (2016) Amygdala-ventral striatum circuit activation decreases long-term fear. eLife 5:e12669.

Cox RW, Chen G, Glen DR, Reynolds RC, Taylor PA (2017) FMRI clustering in AFNI: false-positive rates redux. Brain Connect 7:152–171.

Craske MG, Kircanski K, Zelikowsky M, Mystkowski J, Chowdhury N, Baker A (2008) Optimizing inhibitory learning during exposure therapy. Behav Res Ther 46:5–27.

Craske MG, Treanor M, Conway CC, Zbozinek T, Vervliet B (2014) Maximizing exposure therapy: an inhibitory learning approach. Behav Res Ther 58C:10–23.

Do-Monte FH, Manzano-Nieves G, Quiñones-Laracuente K, Ramos-Medina L, Quirk GJ (2015) Revisiting the role of infralimbic cortex in fear extinction with optogenetics. J Neurosci 35:3607–3615.

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

Dunsmoor JE, Murty VP, Davachi L, Phelps EA (2015b) Emotional learning selectively and retroactively strengthens memories for related events. Nature 520:345–348.

Dunsmoor JE, Campese VD, Ceceli AO, LeDoux JE, Phelps EA (2015c) Novelty-facilitated extinction: providing a novel outcome in place of an expected threat diminishes recovery of defensive responses. Biol Psychiatry 78:203–209.

Dunsmoor JE, Kubota JT, Li J, Coelho CA, Phelps EA (2016) Racial stereotypes impair flexibility of emotional learning. Soc Cogn Affect Neurosci 11:1363–1373.

Dunsmoor JE, Kroes MCW, Moscatelli CM, Evans MD, Davachi L, Phelps EA (2018) Event segmentation protects emotional memories from competing experiences encoded close in time. Nat Hum Behav 2:291–299.

Duvarci S, Paré D (2014) Amygdala microcircuits controlling learned fear. Neuron 82:966–980.

Ehrlich I, Humeau Y, Grenier F, Ciocchi S, Herry C, Lüthi A (2009) Amygdala inhibitory circuits and the control of fear memory. Neuron 62:757–771.

Ekman P, Friesen WV (1976) Measuring facial movement. Environ Psychol Nonverbal Behav 1:56–75.

Foa EB, Kozak MJ (1986) Emotional processing of fear: exposure to corrective information. Psychol Bull 99:20–35.

Foa EB, Steketee G, Rothbaum BO (1989) Behavioral cognitive conceptualizations of post-traumatic stress disorder. Behav Ther 20:155–176.

Frank DW, Dewitt M, Hudgens-Haney M, Schaeffer DJ, Ball BH, Schwarz NF, Hussein AA, Smart LM, Sabatinelli D (2014) Emotion regulation: quantitative meta-analysis of functional activation and deactivation. Neurosci Biobehav Rev 45:202–211.

Friston KJ, Buechel C, Fink GR, Morris J, Rolls E, Dolan RJ (1997) Psychophysiological and modulatory interactions in neuroimaging. Neuroimage 6:218–229.

Fullana MA, Harrison B, Soriano-Mas C, Vervliet B, Cardoner N, Àvila-Parcet A, Radua J (2016) Neural signatures of human fear conditioning: an updated and extended meta-analysis of fMRI studies. Mol Psychiatry 21:500–508.

Fullana MA, Albajes-Eizagirre A, Soriano-Mas C, Vervliet B, Cardoner N, Benet O, Radua J, Harrison BJ (2018a) Amygdala where art thou? Neurosci Biobehav Rev. Advance online publication. Retrieved Jun 7, 2018. doi: 10.1016/j.neubiorev.2018.06.003.

Fullana MA, Albajes-Eizagirre A, Soriano-Mas C, Vervliet B, Cardoner N, Benet O, Radua J, Harrison BJ (2018b) Fear extinction in the human brain: a meta-analysis of fMRI studies in healthy participants. Neurosci Biobehav Rev 88:16–25.

Giustino TF, Maren S (2015) The role of the medial prefrontal cortex in the conditioning and extinction of fear. Front Behav Neurosci 9:298.

Grady AK, Bowen KH, Hyde AT, Totsch SK, Knight DC (2016) Effect of continuous and partial reinforcement on the acquisition and extinction of human conditioned fear. Behav Neurosci 130:36–43.

Green SR, Kragel PA, Fecteau ME, LaBar KS (2014) Development and validation of an unsupervised scoring system (Autonomate) for skin conductance response analysis. Int J Psychophysiol 91:186–193.

Haaker J, Gaburro S, Sah A, Gartmann N, Lonsdorf TB, Meier K, Singewald N, Pape HC, Morellini F, Kalisch R (2013) Single dose of L-dopa makes extinction memories context-independent and prevents the return of fear. Proc Natl Acad Sci U S A 110:E2428–E2436.

Hartley CA, Phelps EA (2010) Changing fear: the neurocircuitry of emotion regulation. Neuropsychopharmacology 35:136–146.

Huff NC, Hernandez JA, Blanding NQ, LaBar KS (2009) Delayed extinction attenuates conditioned fear renewal and spontaneous recovery in humans. Behav Neurosci 123:834–843.

Krabbe S, Gründemann J, Lüthi A (2018) Amygdala inhibitory circuits regulate associative fear conditioning. Biol Psychiatry 83:800–809.

Kroes MC, Tona KD, den Ouden HE, Vogel S, van Wingen GA, Fernández G (2016) How administration of the beta-blocker propranolol before extinction can prevent the return of fear. Neuropsychopharmacology 41: 1569–1578.

Kroes MC, Dunsmoor JE, Lin Q, Evans M, Phelps EA (2017) A reminder before extinction strengthens episodic memory via reconsolidation but fails to disrupt generalized threat responses. Sci Rep 7:10858.

Laborda MA, Miller RR (2012) Reactivated memories compete for expression after pavlovian extinction. Behav Processes 90:20–27.

Laborda MA, McConnell BL, Miller RR (2011) Behavioral techniques to reduce relapse after exposure therapy: applications of studies of experimental extinction. In: Associative learning and conditioning theory: human and non-human applications (Schachtman TR, Reilly S, eds), pp 79–103. New York: Oxford UP.

Laird AR, Robinson JL, McMillan KM, Tordesillas-Gutiérrez D, Moran ST, Gonzales SM, Ray KL, Franklin C, Glahn DC, Fox PT, Lancaster JL (2010) Comparison of the disparity between Talairach and MNI coordinates in functional neuroimaging data: validation of the Lancaster transform. Neuroimage 51:677–683.

Lancaster JL, Woldorff MG, Parsons LM, Liotti M, Freitas CS, Rainey L, Kochunov PV, Nickerson D, Mikiten SA, Fox PT (2000) Automated Talairach atlas labels for functional brain mapping. Hum Brain Mapp 10:120–131.

Larrauri JA, Schmajuk NA (2008) Attentional, associative, and configural mechanisms in extinction. Psychol Rev 115:640–676.

Le Pelley ME (2004) The role of associative history in models of associative learning: a selective review and a hybrid model. Q J Exp Psychol B 57:193–243.

Li J, Schiller D, Schoenbaum G, Phelps EA, Daw ND (2011) Differential roles of human striatum and amygdala in associative learning. Nat Neurosci 14:1250–1252.

Li Y, Nakae K, Ishii S, Naoki H (2016) Uncertainty-dependent extinction of fear memory in an amygdala-mPFC neural circuit model. PLoS Comput Biol 12:e1005099.

Likhtik E, Pelletier JG, Paz R, Paré D (2005) Prefrontal control of the amygdala. J Neurosci 25:7429 –7437.

Lonsdorf TB, Menz MM, Andreatta M, Fullana MA, Golkar A, Haaker J, Heitland I, Hermann A, Kuhn M, Kruse O, Meir Drexler S, Meulders A, Nees F, Pittig A, Richter J, Römer S, Shiban Y, Schmitz A, Straube B, Vervliet B, et al. (2017) Don't fear 'fear conditioning': methodological considerations for the design and analysis of studies on human fear acquisition, extinction, and return of fear. Neurosci Biobehav Rev 77:247–285.

Lucas K, Luck CC, Lipp OV (2018) Novelty-facilitated extinction and the reinstatement of conditional human fear. Behav Res Ther 109:68 –74.

Maldjian JA, Laurienti PJ, Kraft RA, Burdette JH (2003) An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. Neuroimage 19:1233–1239.

Maren S (2014) Nature and causes of the immediate extinction deficit: a brief review. Neurobiol Learn Mem 113:19 –24.

Maren S, Chang CH (2006) Recent fear is resistant to extinction. Proc Natl Acad Sci U S A 103:18020 –18025.

Mechias ML, Etkin A, Kalisch R (2010) A meta-analysis of instructed fear studies: implications for conscious appraisal of threat. Neuroimage 49:1760 –1768.

Milad MR, Quirk GJ (2012) Fear extinction as a model for translational neuroscience: ten years of progress. Annu Rev Psychol 63:129 –151.

Milad MR, Wright CI, Orr SP, Pitman RK, Quirk GJ, Rauch SL (2007) Recall of fear extinction in humans activates the ventromedial prefrontal cortex and hippocampus in concert. Biol Psychiatry 62:446 – 454.

Milad MR, Pitman RK, Ellis CB, Gold AL, Shin LM, Lasko NB, Zeidan MA, Handwerger K, Orr SP, Rauch SL (2009) Neurobiological basis of failure to recall extinction memory in posttraumatic stress disorder. Biol Psychiatry 66:1075–1082.

Milad MR, Vidal-Gonzalez I, Quirk G (2004) Electrical stimulation of medial prefrontal cortex reduces conditioned fear in a temporally specific manner. Behav Neurosci 118:389 –394.

Miller RR, Laborda MA (2011) Preventing recovery from extinction and relapse: a product of current retrieval cues and memory strengths. Curr Direct Psychol Sci 20:325–329.

Morriss J, Christakou A, van Reekum CM (2016) Nothing is safe: intolerance of uncertainty is associated with compromised fear extinction learning. Biol Psychol 121:187–193.

Myers KM, Davis M (2002) Behavioral and neural analysis of extinction. Neuron 36:567–584.

Ohman A, Mineka S (2001) Fears, phobias, and preparedness: toward an evolved module of fear and fear learning. Psychol Rev 108:483–522.

Ongür D, Ferry AT, Price JL (2003) Architectonic subdivision of the human orbital and medial prefrontal cortex. J Comp Neurol 460:425– 449.

Pape HC, Paré D (2010) Plastic synaptic networks of the amygdala for the acquisition, expression, and extinction of conditioned fear. Physiol Rev 90:419 – 463.

Pavlov IP (1927) Conditioned reflexes. London: Oxford UP.

Pearce JM, Hall G (1980) A model for pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. Psychol Rev 87:532–552.

Phelps EA, Delgado MR, Nearing KI, LeDoux JE (2004) Extinction learning in humans: role of the amygdala and vmPFC. Neuron 43:897–905.

Quirk GJ, Mueller D (2008) Neural mechanisms of extinction learning and retrieval. Neuropsychopharmacology 33:56 –72.

Quirk GJ, Likhtik E, Pelletier JG, Paré D (2003) Stimulation of medial prefrontal cortex decreases the responsiveness of central amygdala output neurons. J Neurosci 23:8800 – 8807.

Raio CM, Hartley CA, Orederu TA, Li J, Phelps EA (2017) Stress attenuates the flexible updating of aversive value. Proc Natl Acad Sci U S A 114:11241–11246.

Rescorla RA (2004) Spontaneous recovery. Learn Mem 11:501–509.

Rescorla RA, Wagner AR (1972) A theory of Pavlovian conditioning: variations in the effectiveness of reinforcement and nonreinforcement: New York: Appleton-Century-Crofts.

Schiller D, Cain CK, Curley NG, Schwartz JS, Stern SA, Ledoux JE, Phelps EA (2008) Evidence for recovery of fear following immediate extinction in rats and humans. Learn Mem 15:394 – 402.

Schiller D, Monfils MH, Raio CM, Johnson DC, Ledoux JE, Phelps EA (2010) Preventing the return of fear in humans using reconsolidation update mechanisms. Nature 463:49 –53.

Schiller D, Kanen JW, LeDoux JE, Monfils MH, Phelps EA (2013) Extinction during reconsolidation of threat memory diminishes prefrontal cortex involvement. Proc Natl Acad Sci U S A 110:20040 –20045.

Sehlmeyer C, Schoning S, Zwitserlood P, Pfleiderer B, Kircher T, Arolt V, Konrad C (2009) Human fear conditioning and extinction in neuroimaging: a systematic review. PLoS One 4:e5865.

Senn V, Wolff SB, Herry C, Grenier F, Ehrlich I, Gründemann J, Fadok JP, Müller C, Letzkus JJ, Lüthi A (2014) Long-range connectivity defines behavioral specificity of amygdala neurons. Neuron 81:428 – 437.

Sierra-Mercado D, Padilla-Coreano N, Quirk GJ (2011) Dissociable roles of prelimbic and infralimbic cortices, ventral hippocampus, and basolateral amygdala in the expression and extinction of conditioned fear. Neuropsychopharmacology 36:529 –538.

Solomon RL, Wynne LC (1954) Traumatic avoidance learning: the principles of anxiety conservation and partial irreversibility. Psychol Rev 61:353–385.

Strobel C, Marek R, Gooch HM, Sullivan RK, Sah P (2015) Prefrontal and auditory input to intercalated neurons of the amygdala. Cell Rep 10:1435–1442.

Sutton RS (1992) Adapting bias by gradient descent: An incremental version of δ-bar-δ. In: Proceedings of the Tenth National Conference on Artificial Intelligence (Swarout WR, ed), pp 171–176. Menlo Park, CA: American Association for Artificial Intelligence.

Tovote P, Fadok JP, Lüthi A (2015) Neuronal circuits for fear and anxiety. Nat Rev Neurosci 16:317–331.

Vervliet B, Craske MG, Hermans D (2013) Fear extinction and relapse: state of the art. Annu Rev Clin Psychol 9:215–248.

Wagner AR (1981) SOP: A model of automatic memory processing in animal behavior. In: Information processing in animals: memory mechanisms (Spear NE, Miller RR, eds), pp 5– 47. Hillsdale, NJ: Erlbaum.