Cognitive, Affective, & Behavioral Neuroscience (2020) 20:983–997
https://doi.org/10.3758/s13415-020-00814-4

Extinction learning alters the neural representation of conditioned fear

John L. Graner[1] · Daniel Stjepanović[1,2] · Kevin S. LaBar[1]

Published online: 27 July 2020

Abstract

Extinction learning is a primary means by which conditioned associations to threats are controlled and is a model system for emotion dysregulation in anxiety disorders. Recent work has called for new approaches to track extinction-related changes in conditioned stimulus (CS) representations. We applied a multivariate analysis to previously -collected functional magnetic resonance imaging data on extinction learning, in which healthy young adult participants (N = 43; 21 males, 22 females) encountered dynamic snake and spider CSs while passively navigating 3D virtual environments. We used representational similarity analysis to compare voxel-wise activation t-statistic maps for the shock-reinforced CS (CS+) from the late phase of fear acquisition to the early and late phases of extinction learning within subjects. These patterns became more dissimilar from early to late extinction in *a priori* regions of interest: subgenual and dorsal anterior cingulate gyrus, amygdala and hippocampus. A whole-brain searchlight analysis revealed similar findings in the insula, mid-cingulate cortex, ventrolateral prefrontal cortex, somatosensory cortex, cerebellum, and visual cortex. High state anxiety attenuated extinction-related changes to the CS+ patterning in the amygdala, which suggests an enduring threat representation. None of these effects generalized to an unreinforced control cue, nor were they evident in traditional univariate analyses. Our approach extends previous neuroimaging work by emphasizing how evoked neural patterns change from late acquisition through phases of extinction learning, including those in brain regions not traditionally implicated in animal models. Finally, the findings provide additional support for a role of the amygdala in anxiety-related persistence of conditioned fears.

Keywords Extinction learning · Representational similarity analysis · Functional MRI · Fear conditioning

The ability to suppress fear responses when they are no longer appropriate is a hallmark of healthy emotion regulation. Extinction learning provides a powerful means to override acquired fears by repeatedly exposing individuals to threat encounters in a safe context without aversive consequences. Extinction learning is thought to update the affective representation of a threat in a context-dependent way such that a new, safe memory competes with the original fear acquisition memory and diminishes defensive reactions (Bouton, 1993) (but see Dunsmoor et al., 2015). Given that extinction learning forms the basis of exposure-based therapies, which are effective in treating specific phobias and posttraumatic stress disorder (PTSD), there is much interest in understanding its neural basis (Marks & Tobena, 1990).

Neurobiological models, derived largely from rodent studies, have implicated inhibitory connections between the ventromedial prefrontal cortex (vmPFC) and the amygdala as being critical for extinction learning. The vmPFC engages inhibitory pathways within the basolateral and centromedial complexes of the amygdala that consolidate memories of the extinction training experience and reduce subsequent conditioned fear responses (Milad & Quirk, 2012). Hippocampal input can up- or down-regulate activity in these structures, which contributes to the context-specificity of extinction learning and susceptibility to relapse following context shifts

This work was supported by NSF grant BCS 1460909 to K.S.L. We thank Fredrik Åhs for his contributions to experimental design, data collection, and analysis of the univariate results that provided the foundation of the present manuscript.

Electronic supplementary material The online version of this article (https://doi.org/10.3758/s13415-020-00814-4) contains supplementary material, which is available to authorized users.

✉ Kevin S. LaBar
klabar@duke.edu

[1] Center for Cognitive Neuroscience, Duke University, Box 90999, Durham, NC 27708-0999, USA

[2] Present address: Centre for Youth Substance Abuse Research, University of Queensland, St Lucia, Australia

(Maren, 2011). Pharmacologic and electrophysiologic manipulation of activity within this hippocampal-vmPFC-amygdala circuit impacts the recall of extinction learning (Ji & Maren, 2007; Laurent & Westbrook, 2009; Sierra-Mercado et al., 2011), with implications for developing neuromodulatory interventions to treat extinction-resistant fear memories.

Efforts to translate these findings to humans have yielded mixed evidence (Sevenster et al., 2018). A key tenet of the findings from the rodent literature is that plasticity in this core circuit during extinction learning—especially the inhibitory interactions between the vmPFC and amygdala—is critical for establishing the memory trace associating the previously conditioned stimulus (CS) with safety. Thus, neuroimaging studies should show engagement of these structures during extinction learning, and this activity and/or functional connectivity should predict subsequent extinction recall. One challenge to this logic is that both inhibitory and excitatory influences can yield increases in blood-oxygenated-level-dependent (BOLD) signal (Xu, 2015), making it unclear how extinction learning would translate quantitatively into changes in the aggregate BOLD signal. Further complicating matters, the vmPFC tends to exhibit decreases in BOLD signal when engaged by a cognitive task, relative to a resting baseline, as part of the brain's default mode network (Raichle et al., 2001). Despite these challenges, some initial neuroimaging evidence supported a role for the hippocampal-vmPFC-amygdala circuit in extinction processes. For instance, LaBar et al. (1998) showed that the amygdala exhibited a transient response to a shock-reinforced CS (CS+) during early extinction training that declined over subsequent trials. Phelps et al. (2004) found that amygdala activity to a CS+ decreased from acquisition to extinction training, and the amount of extinction learning evident in skin conductance responses (SCRs) predicted vmPFC activity to a CS+ during a subsequent extinction recall session. Milad et al. (2007) reported enhanced vmPFC signaling during recall of an extinguished CS+ along with greater functional coupling between the vmPFC and the hippocampus. Using structural equation modeling, Åhs et al. (2015) showed that the vmPFC mediates amygdala-hippocampal functional coupling to support extinction recall.

Nonetheless, in a recent meta-analysis of 31 human fMRI studies of extinction learning, Fullana et al. (2018) did not find reliable evidence for vmPFC engagement or amygdala down-regulation to a CS+, as would be predicted from the rodent models and this earlier neuroimaging work. Instead, brain regions active during fear acquisition, such as the dACC and insula, consistently contributed to fear extinction as well, and a direct comparison of brain regions that were more engaged during extinction learning than acquisition training yielded no significant effects. Although some evidence of hippocampal and vmPFC engagement during a delayed extinction recall test was found in the meta-analysis, this evidence was restricted to a small number of studies that utilized a particular stimulus comparison (recall for a previously extinguished CS+ compared with recall for an unextinguished CS+). Regardless, this latter finding is inconsistent with some rodent electrophysiological and optogenetic studies showing that vmPFC-amygdala interactions, while important to consolidate extinction memories at the time of initial learning, are not critical for expressing this prior learning at a delayed recall test (Bukalo et al., 2015; Do-Monte et al., 2015).

While null results are challenging to interpret, the Fullana et al. (2018) meta-analysis raises issues regarding possible species differences in the neural processing of extinction and/or methodological limitations of existing neuroimaging studies. One possibility implicated by this meta-analysis is that residual signaling of the threat value of the CS+ may carry over into extinction training and conflate interpretation of fMRI signal changes during extinction. Bolstering this possibility are results from electrophysiological and immunohistochemical studies in rodents showing that residual conditioned fear and extinction representations can be interdigitated (and perhaps compete for expression) by neighboring neuronal pools within the same brain region (for review, see Courtin et al., 2013; Dejean et al., 2015). Thus, more refined pattern representation approaches and/or higher-resolution fMRI studies may be needed to disambiguate threat and safe memory representations during extinction learning.

We addressed this issue by adapting multivariate pattern analysis (MVPA) methods to track extinction-induced changes in the neural signaling of conditioned stimuli. We aimed to characterize the extent to which extinction learning creates a unique patterned representation of the CS relative to that at the end of acquisition training. MVPA considers the contributions of subthreshold voxels rather than relying on the peak signal change in a region, thus obtaining a more comprehensive view of the distributed nature of neural signaling in a brain region associated with a stimulus. These methods can even differentiate voxelwise patterns associated with different functions when the mean signal in an ROI is the same across conditions (for a review of MVPA and its applications, see Kriegeskorte, 2011; Tong & Pratte, 2012). For these reasons, we feel that MVPA may be particularly useful for distinguishing a change in the CS representation from acquisition to extinction training or other kinds of context shifts.

A few studies have investigated the change in the representational similarity of CS+ and CS− stimuli during fear acquisition or extinction learning using MVPA (Visser et al., 2011; Visser et al., 2013; Visser et al., 2015; Visser et al., 2016). The general idea tested in these studies is that as conditioned learning progresses, the CS representation should become more stable across trials. Results from these studies indicate increasing trial-to-trial similarity in the CS+ neural representation through acquisition. This increased similarity was still present upon re-exposure to the CS+ days to weeks later. This maintained similarity was not found for the CS−, which the authors

suggested reflected a more refined (i.e., more specific and reproducible) neural response to the CS+, likely driven by a learned threat association. In all the studies by Visser and colleagues, acquisition and extinction were performed in separate neuroimaging sessions, and there was no direct comparison between the neural patterns across phases of conditioning. This leaves open the question of how and when changes in reinforcement contingencies, such as the transition from acquisition to extinction training, or during reversal learning, alters the specific neural pattern response to the CS that is established through conditioning.

Studies from a separate literature on episodic event segmentation provide guidance for considering how representations of CSs might change across phases of learning. According to event segmentation theory (Zacks et al., 2007), when naturalistic events unfold over time, event boundary markers are established by prediction errors created when incidents deviate from what is expected based on contextual information. This binding of event segments at a boundary facilitates memory encoding, enables separate memory traces for sequential event segments to be established, and refocuses attention to update and align working memory contents to a model of the current event structure (Radvansky & Zacks, 2017). The hippocampus contributes to memory formation based on binding of information at an event boundary (Ben-Yakov et al., 2014; Baldassano et al., 2017), and many cortical brain regions establish new patterns of activity soon after an event transition from one context to another in order to stably signal the new event segment (Baldassano et al., 2017).

As applied to conditioning paradigms, event segmentation theory would predict separable representations of both the CS+ and the CS− from acquisition to extinction training, given that the initiation of extinction involves a novel learning context due to the removal of the unconditioned stimulus (US). Separability should be further enhanced when acquisition and extinction phases are conducted in two distinct spatial environments, which would require binding both the CS+ and CS− to distinct contextual cues across training phases. Nonetheless, if the threat value of the CS+ is initially resistant to change due to carryover of conditioned fear associations from the acquisition phase (as suggested by the re-exposure findings of Visser and colleagues, among others), then its representation may undergo a more delayed shift than that of the CS- at the event boundary. Although it has been hypothesized from associative learning principles that acquisition and extinction training establish distinct engrams to conditioned stimuli (Bouton, 1993), it has been difficult to track these representational changes with current neuroimaging methods.

In the current study, we re-analyzed fMRI data collected from a prior fear conditioning experiment (Åhs et al., 2015) in which acquisition and extinction data were collected in the same imaging session, allowing representational similarity analysis (RSA) to be performed between the two phases without introducing day or session effects in the pattern similarity metrics. The conditioning paradigm consisted of a 3-D virtual reality (VR) environment in which participants encountered conditioned stimuli (snakes/spiders) during passive navigation. Conditioned fear acquisition and extinction learning took place in different immersive VR contexts (an indoor scene and an outdoor scene, counterbalanced across participants), which provided a clear event boundary to differentiate memory engrams of each phase for both the CS+ and CS− stimuli. We previously reported that this paradigm elicits changes in functional connectivity 24 hours after extinction learning such that vmPFC gating of amygdala-hippocampal connectivity promoted extinction recall whereas dACC gating of this connectivity promoted fear renewal (Åhs et al., 2015). However, consistent with the meta-analytic results of Fullana et al. (2018), we failed to find significant signal changes in these regions during the initial extinction learning itself using a traditional univariate analysis, with the exception of dACC activity in the CS+ > CS − contrast during early extinction. We applied RSA to test the hypothesis, based on the existing event boundary and extinction meta-analysis literatures, that the conditioned fear representation of the CS+ changes from early to late training blocks of extinction learning and that this change would be larger relative to that of an unreinforced stimulus (CS−). We also tested the related hypothesis that the neural representation of the CS+ would undergo less change when initially moving from the acquisition context to the novel extinction context than would the representation of the CS−, due to the carry-over associations acquired through the fear-conditioning process. Given that representations of conditioned stimuli involve cell assemblies with coordinated firing patterns that span several frontolimbic structures (Courtin et al., 2013; Rozeske & Herry, 2018), we did not have *a priori* hypotheses about the specificity of these effects within the core fear-conditioning network.

Because extinction learning is the fundamental component of exposure-based therapies, it is important to understand how individual differences in anxiety impact the ability of extinction to modify threat representations. Anxiety disorders are associated with aberrant learning processes following fear conditioning as individuals overgeneralize or fail to regulate their affective responses (Lissek, 2012). Thus, high-anxious individuals should exhibit less change in the representation of the CS+ as extinction training progresses as a result of its lingering association with threat. We tested this hypothesis in a core fear conditioning network, including the amygdala, hippocampus, dACC, and vmPFC.

Method

Participants

Healthy, right-handed young adults (N = 45) were presented with dynamic snake and spider conditioned stimuli while passively navigating 3D virtual environments using stereoscopic video projection. Two participants were unable to complete a portion of the original data collection and were excluded from analysis. The final sample size was N = 43 (21 males, 22 females; mean age = 28.7 years). Participants were compensated $20/hr for participating and provided written, informed consent consistent with procedures approved by the Duke University Medical Center Institutional Review Board. The initial univariate and structural equation modeling data analyses from this study, along with the behavioral and psychophysiological data, were previously reported in Åhs et al. (2015); here we report new results from a follow-up MVPA analysis.

Stimuli

Fear acquisition and extinction occurred in two different 3D virtual environment contexts, an indoor apartment scene and an outdoor woods scene, counterbalanced across participants. Assignment of CS type (CS+/CS−) to stimulus type (snake/spider) also was counterbalanced across participants. The 3D conditioned stimuli and virtual environments were constructed in Maya (Autodesk, San Rafael, CA). These stimuli were then presented in Virtools (Dassault Systeme, Paris, France) via MRI-compatible goggles (VisuaStim, Resonance Technology Inc., Northridge, CA). During passive navigation, participants viewed the virtual environments from a first-person perspective as if they were walking through them. During presentation of CSs, this progression was paused while the given CS was rendered into the scene for 4 sec. The interstimulus interval for both acquisition and extinction was jittered in a range between 10 and 14 sec.

Presentation of the unconditioned stimulus (US) overlapped the final 16 ms of paired CS+ events during the fear acquisition phase. The US was an unpleasant shock produced by an MP-150 BIOPAC system (STM-100 and STM-200 modules, BIOPAC systems, Goleta, CA) and delivered through electrodes (EL507, BIOPAC systems) attached to the right wrist. The magnitude of the shock was set for each individual participant during the experimental setup using an ascending staircase procedure and was calibrated using the participant's feedback to be "highly annoying but not painful" (mean and standard deviation of US voltage = 49 ± 17 V). US presentations were associated with the same CS+ presentations for each participant and were spaced throughout acquisition training using a pseudorandomization procedure.

Fear conditioning task

The conditioning procedure consisted of a habituation phase, an acquisition phase, and an extinction phase. Participants were instructed to predict when a snake or spider would be paired with a shock. Participants responded "no," "unsure," or "yes" in response to each CS via button press on a MRI-compatible button box. CS presentation was ordered pseudorandomly such that no more than two stimuli of the same type appeared consecutively. Participants returned the next day for a fear renewal test, but only data from Day 1 (fear acquisition and extinction) are analyzed here.

The habituation phase consisted of 4 CS+ and 4 CS− presentations without reinforcement. This phase was included only to reduce initial orienting responses to the stimuli and to acclimate the participants to the immersive VR environment; no data were analyzed from this phase. The fear acquisition phase included 16 CS+ and 16 CS− presentations. A partial reinforcement paradigm was used during this phase, with 5 of the 16 CS+'s (31%) being paired with the US. The extinction phase also contained 16 CS+ and 16 CS- presentations, but none of the stimuli were paired with the US. A brief pause separated each of the three study phases. During these interphase periods, participants reported their current level of anxiety on a scale from 1 ("not anxious at all") to 10 ("worst imaginable anxiety").

Magnetic resonance image collection

Functional and anatomical brain images were acquired on a General Electric Signa EXCITE HD 3.0 Tesla magnetic resonance imaging (MRI) scanner with 40-mT/m gradients using an 8-channel head coil (General Electric, Waukesha, WI). Before functional imaging, a high-resolution, T1-weighted image was collected with a 3D fast Spoiled Gradient Echo sequence (repetition time (TR) = 500 ms; echo time (TE) = 31 ms; image matrix = 256 x 256; 68 contiguous slices; voxel size = 0.9375 x 0.9375 x 1.9 mm). Functional images were collected using a SENSE™ spiral-in sequence (acquisition matrix = 64 x 64; field of view = 256 x 256; flip angle = 60°; 34 slices; interleaved slice acquisition; slice thickness = 3.8 mm; no slice gaps; TR = 2,000 ms; TE = 27 ms).

Skin conductance response collection and data processing

Skin conductance responses (SCR) were recorded continuously during acquisition and extinction using an MP-150 BIOPAC system through MRI-compatible Ag/AgCl electrodes on the palmar surface of participants' left hands and analyzed using Autonomate software (Green, et al., 2014). SCR data validated the experimental manipulation, as previously reported in Åhs et al. (2015). We reanalyzed the SCR

data by normalizing each stimulus presentation's response to the maximal response to the unconditioned stimulus for each subject ($SCR_{final} = SCR/MAX_{US}$) to achieve the goal of comparing standardized SCR and fMRI responses within-subjects.

Average SCR values for early and late extinction were calculated for each participant. "Early" and "Late" phases of extinction refer to, respectively, the first 8 events of each type and the last 8 events of each type (see "Imaging Data Processing" section below for more information on this event binning). These values were analyzed using a repeated measures ANOVA with factors of CS Type (CS+/CS−) and Time (Early/Late extinction blocks).

For completeness, the normalized SCR and reported Shock Expectancy values for each trial are reported in the Supplemental Material (Figures S1 and S2).

Shock expectancy data processing

Shock expectancy responses of "no," "unsure," and "yes" were coded as 0, 0.5, and 1.0, respectively. Average expectancy values for Early and Late Extinction were calculated for each participant. These values were analyzed using a repeated-measures ANOVA with factors of CS Type (CS+/CS−) and Time (Early/Late extinction blocks).

Imaging data processing

Preprocessing of the raw imaging data for the multivariate analysis was performed from scratch in AFNI (Cox, 1996). The primary difference between the preprocessing performed in this study and the preprocessing performed by Åhs et al. (2015) previously is the exclusion of spatial smoothing in the current pipeline. First, the anatomical data were skull-stripped and separately registered to the functional data and warped to MNI-152 space (using the 2 mm^3 template as a target). This procedure created two sets of transformation files: one converting between anatomical space and functional space, and the other converting between anatomical space and MNI space. The functional data underwent slice-time correction, motion-correction, and warping into standard MNI-152 space. The warp to standard space was performed by applying the inversion of the transform from anatomical space to functional space and the transform from anatomical space to MNI space. Additionally, the target grid of the functional data in the warp to standard space was set to 3.5 mm^3. This step was done to minimize the effects of interpolation and resampling on the functional data. In order to better facilitate multivariate analysis, no smoothing was applied to the functional image data. The standard-space anatomical images were then segmented to create grey matter masks for each participant. These binary masks were averaged together, and voxels with values greater than 0.5 (i.e., at least half the participants' individual masks contained the voxel) were retained to create a final group gray matter mask.

Creation of first-level statistical maps for use in the multivariate analysis was performed using SPM8 (Wellcome Trust Centre for Neuroimaging, University College London, London, UK). T-statistic maps were generated for each participant using a general linear model (GLM) approach for the following conditions: CS+ and CS− responses in Early Acquisition, Late Acquisition, Early Extinction, and Late Extinction. "Early" and "Late" refer to the same epochs as described above for the SCR data processing. Standard regressor creation was used for the GLM, convolving a Gaussian hemodynamic response function with boxcar functions based on the onset times and durations of the presentations of each stimulus type. Thus, the first-level GLM produced 8 t-statistic maps (2 CS types x 4 time epochs). Six of these maps (Late Acquisition, Early Extinction, and Late Extinction for each CS type) were used as inputs to the multivariate representational similarity analysis (RSA).

We had originally explored performing a trial-by-trial analysis of these data, similar to the analyses performed by Visser et al. (2013, 2015). However, the task protocol was not originally designed for such an analysis, which requires fixed times of sufficient duration between consecutive trials of each condition type to limit confounds due to intrinsic noise correlations. Fixed trial intervals are typically not implemented in fear conditioning paradigms to minimize temporal confounds in CS onset predictability. Thus, we expected that the trial-to-trial dissimilarity metrics would be highly impacted by temporal autocorrelation inherent in BOLD data. A preliminary event-wise RSA showed that, as anticipated, there was a significant relationship between trial-to-trial dissimilarity and trial-to-trial time separation in both acquisition and extinction for both CS+ and CS− events (see *Supplemental Materials* for details). Given this methodological confound, we abandoned the trial-by-trial analysis in favor of modeling groups of events across Early and Late time epochs of each phase. This event grouping had been used in the univariate analysis performed in the original study (Åhs et al., 2015); it provides a more stable estimate of the CS+ and CS− representations during Late Acquisition relative to a single-trial estimate, which serves as the comparison to the extinction phases; and it allows each bin to contain a similar number of trials across participants. Given that the CS+/CS− discrimination happened relatively quickly in both the shock expectancy and SCR measures (as is typical of human fear conditioning studies; see *Supplementary Materials*), this trial block parsing roughly corresponds to an early active learning phase and a later learning maintenance phase.

The preliminary trial-to-trial RSA also showed that CS+ events paired with the US had response patterns that were different from those of the other CS+ events (see *Supplemental Materials* for details on this analysis). Because

responses to paired events could include US processing as a potential confound, the multivariate analysis focused on the unpaired CS+ trials, with the paired CS+ trials modeled as a separate regressor in the first-level GLM (note: the binning of events into "Early" and "Late" epochs was done counting all CS+ events and did not change following the creation of the paired CS+ regressor). Treatment of US-paired CS+ events as a condition of no interest has also been done in previous fear conditioning fMRI studies (Visser et al., 2011; Sehlmeyer et al., 2011). A regressor was also included for US events as well as the six motion correction parameters (translation and rotation in x, y, and z) for each TR. Finally, TR-wise motion-censoring regressors were created and included in the model in cases where significant motion was present. Specifically, a vector was created for each pair of TRs, consisting of the difference in the six motion correction parameters between the two time points. A motion-censor regressor was created for a TR if the Euclidean norm of the vector associated with the TR and the preceding TR was greater than 0.4. These motion-censoring regressors contained a value of "1" at the TR to be censored and a value of "0" for every other TR in order to minimize the influence of excess motion on the GLM beta estimates for the regressors of interest. The mean percent of motion-censored TRs across all Acquisition images was 1.9% and ranged from 0% to 9.5%. The mean percent of motion-censored TRs across all Extinction images was 1.4% and ranged from 0% to 13.7%.

The analysis of univariate data (described below) was performed on the first-level results previously reported by Åhs et al. (2015). No new preprocessing was performed for that analysis.

Regions of interest

The *a priori* regions of interest (ROIs) defined in Åhs et al. (2015) were again used here after being resampled to match the voxel size of the newly processed functional data. These ROIs were created using selections of the Wake Forest University PickAtlas software (Maldjian et al., 2003) and included amygdala, hippocampus, dorsal anterior cingulate cortex (dACC), and ventral medial prefrontal cortex (vmPFC). The AAL left and right hippocampus ROIs were split into two parts, anterior and posterior, by bisecting them at y = −24 mm in MNI space (Poppenk et al., 2013). The amygdala ROIs were taken from the TD library (Maldjian et al., 2003). The dACC ROI used was the portion of the AAL library region superior to the genu. The vmPFC region was created by dilating the TD library BA 25 ROI by 2 mm. For the current analysis, the left and right ROIs for the anterior hippocampus, posterior hippocampus, and amygdala were combined for each region to create bilateral ROIs. Although only results for the posterior hippocampus ROI were reported in the previous analysis (Åhs et al., 2015), both the anterior and posterior portions were included in the current multivariate analysis for completeness. Thus, there are five ROIs used in the multivariate analysis (bilateral anterior hippocampus, bilateral posterior hippocampus, bilateral amygdala, vmPFC, and dACC) and four ROIs reported in the univariate analysis (bilateral posterior hippocampus, bilateral amygdala, vmPFC, and dACC).

Representational similarity analysis of functional imaging data

Representational Similarity Analysis (RSA) was performed using the rsatoolbox (https://github.com/rsagroup/rsatoolbox) written for the Matlab programming environment (The Mathworks, Inc., Natick, MA). Inputs to the RSA were the t-statistic maps produced by the first-level analysis described above. The purpose of the RSA was to produce metrics estimating the degree of neural pattern dissimilarity between Late Acquisition and each of the two time epochs (Early and Late) of Extinction. The general process for the calculation of a dissimilarity metric between two conditions (e.g., Late Acquisition CS+ and Early Extinction CS+) in a given ROI for a single participant's fMRI data set is as follows: 1) Extract the ROI voxel values from the participant's first-level t-statistic map from the first condition (e.g., Late Acquisition, CS+); 2) Create a single row vector of these values; 3) Repeat steps 1 and 2 for the second condition (e.g., Early Extinction, CS+); 4) Calculate the correlation between these two vectors; 5) Dissimilarity metric between the two conditions for the given ROI for the given participant is 1 minus this correlation value. Thus, if the spatial patterns of t-statistics for the two conditions are similar, the calculated vector correlation will be high and the dissimilarity metric to be relatively low. On the other hand, if the spatial t-statistic patterns have little in common between the two conditions, the calculated vector correlation will be low, and thus, the dissimilarity metric will be relatively high.

One set of dissimilarity metrics was calculated comparing patterns in Late Acquisition and Early Extinction ("Early Dissimilarity" in Fig. 1). A second set of dissimilarity metrics was calculated comparing patterns in Late Acquisition and Late Extinction ("Late Dissimilarity" in Fig. 1). Both of these sets were created for each of the two stimulus types (CS+ and CS−) for each of the five *a priori* ROIs for each participant. These dissimilarity metrics were used to test our hypotheses regarding the neural representations of the stimuli through extinction: the representation of the CS+ would change through extinction, shown by greater dissimilarity between Late Acquisition and Late Extinction than between Late Acquisition and Early Extinction; this change in the representation of the CS+ would be larger than the change in the representation of the CS−, shown by a larger change in dissimilarity metrics for the CS+ than for the CS− between Early

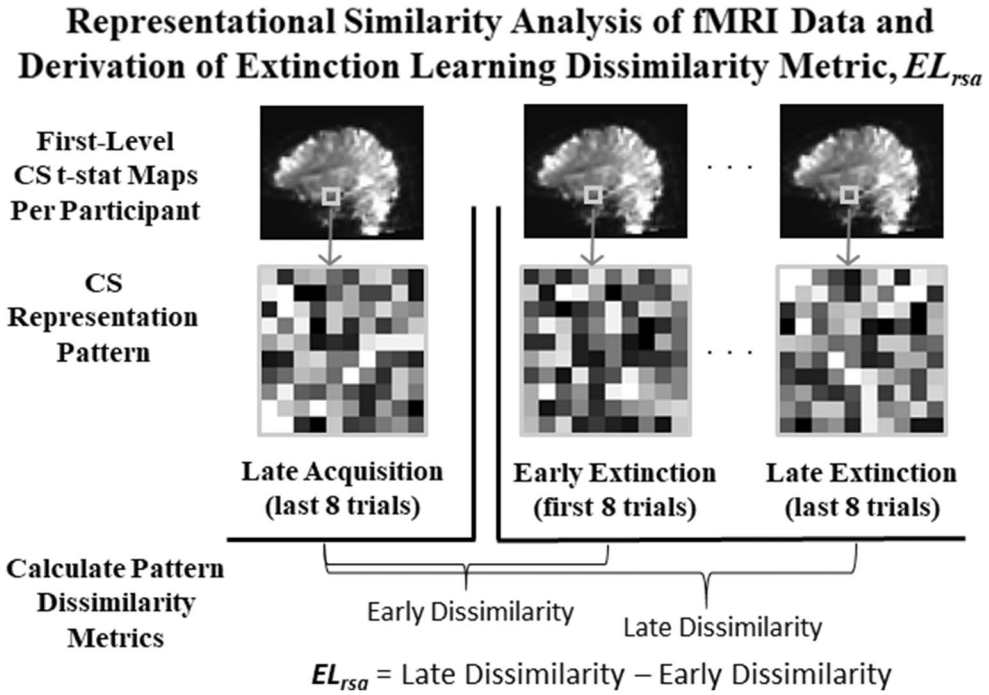

Fig. 1 Derivation of a multivariate neural metric of change in representational dissimilarity across extinction learning (EL_{rsa}). Within-subject changes in the similarity of the reinforced stimulus (CS+) representation are calculated voxel-wise from Late Acquisition to Early Extinction (*Early Dissimilarity*) and from Late Acquisition to Late Extinction (*Late Dissimilarity*) and then compared to each other. This procedure is repeated using the nonreinforced stimulus (CS−) trials to determine the specificity of the extinction-induced change in neural representation

and Late Extinction; and the representation of the CS+ would undergo less change across the event boundary between acquisition and extinction, as shown by a greater dissimilarity metric between Late Acquisition and Early Extinction in the CS− than in the CS+. These hypotheses were tested in each of the five ROIs independently. This procedure was done by entering the dissimilarity metrics into a multi-factor analysis of variance (ANOVA) with CS Type (CS+, CS−), Time (Early Extinction, Late Extinction), and ROI as main factors and all two-way and three-way interaction terms included. Follow-up *t*-tests were performed to further investigate the relationships driving significant model terms.

In response to reviewer requests, we also created average dissimilarity metrics between CS+ and CS− events within Early Acquisition, Late Acquisition, Early Extinction, and Late Extinction. These results are presented and discussed in Supplemental Figure S6.

RSA metric across extinction learning (EL_{rsa})

A participant-wise metric summarizing the change in CS+ representational dissimilarity across Extinction Learning, EL_{rsa+}, was created based on the voxel-wise neural representation data. EL_{rsa+} is the difference between "Late Dissimilarity" and "Early Dissimilarity" for the CS+ (Fig. 1). The same RSA metric was also computed for the CS- (EL_{rsa-}). To determine whether the extinction-induced change in representation of the CS+ was related to individual differences in conditioned learning, we regressed the EL_{rsa+} against the SCR index of differential conditioning (normalized mean Late Acquisition SCR ($(SCR_{CS+} - SCR_{CS-})/MAX_{US}$)) across participants. We also regressed EL_{rsa+} against self-reported anxiety following Late Acquisition and against self-reported change in shock expectancy during Extinction ($ShockExp_{CS+,Early} - ShockExp_{CS+,Late}$). All regressions performed with EL_{rsa+} were also performed with EL_{rsa-} Regressions were performed by using the IBM SPSS Statistics software.

Whole-brain searchlight RSA

To complement the ROI-based RSA described above, an exploratory whole-brain searchlight analysis was performed to look for regions of the brain, in addition to our *a priori* ROIs, that showed changes in local neural patterns, again relative to Late Acquisition, between the two time epochs of extinction. This analysis was performed as a data-driven approach to identify other brain regions that may potentially be involved in or influenced by extinction learning. Dissimilarity metrics were calculated as described above for the ROI-based RSA. However, rather than using anatomically based ROIs, each

voxel in the brain had a 7-mm radius sphere ROI centered on it. Dissimilarity metrics calculated from a given voxel's sphere ROI were then assigned to that voxel, producing whole-brain dissimilarity metric maps. Two such whole-brain dissimilarity maps were created for each participant: one for Late Acquisition-to-Early Extinction dissimilarity, and one for Late Acquisition-to-Late Extinction dissimilarity, both using the CS+. The group gray matter mask was applied to these maps and a paired t-test was performed on each voxel (AFNI's 3dttest++; paired within participant; including whole-brain maps from all participants) to determine whether there was significant difference between the two dissimilarity metrics in that voxel.

An identical searchlight analysis was performed for the CS − as well, again exploring differences in Late Acquisition-to-Late Extinction dissimilarity and Late Acquisition-to-Early Extinction dissimilarity.

Searchlight results were corrected for multiple comparisons using a family-wise error approach (using 3dttest++'s -Clustsim option). Ten thousand random permutations of the *t*-test residuals were performed to simulate the null-condition t-distribution. This procedure allows the estimation of the number of clusters of various sizes that would randomly occur based solely on noise at a range of voxel-wise *t*-test *p*-values without assuming a specific structure of the noise spatial autocorrelation. At a voxel-wise *t*-test *p*-value of 0.001, a cluster size of at least 14 voxels (CS+) or 15 voxels (CS-) corresponded to a cluster-wise corrected *p*-value (alpha) less than 0.05. Generation of the null-condition distribution and the t-test results were spatially restricted to the group gray-matter mask created from the anatomical data.

Comparison to standard univariate extinction metrics

To determine if the multivariate metrics developed here contributed additional value beyond that of the univariate data, we put the previously estimated univariate results (reported in Table S1, Åhs et al. (2015)) through the same analyses as the new multivariate metrics. For each participant, for each CS type, and for each ROI, the average Late Acquisition activation value was subtracted from each of the Early Extinction and Late Extinction values. These univariate metrics were entered into a multi-factor ANOVA with CS Type (CS+, CS−), Time (Early Extinction, Late Extinction), and ROI as factors. A mean variable was also created for each participant that represented the change in activation to the CS+ from Early to Late Extinction (Late Extinction CS+ - Early Extinction CS+). This contrast was treated as a univariate equivalent of EL_{rsa+} and was similarly regressed against SCR indices of differential conditioned learning and reported anxiety across participants using SPSS. Individual univariate regressions were carried out for each of SCR and reported anxiety for each ROI, and the final significance values were Holm-corrected for multiple comparisons based on the number of ROIs (Holm, 1979).

Results

Neural representation of the CS+ becomes more dissimilar from late acquisition as extinction progresses

The representational dissimilarity metrics between Late Acquisition and Early Extinction trials (Early Dissimilarity), and between Late Acquisition and Late Extinction trials (Late Dissimilarity), are shown in Fig. 2A for each CS type. Inspection of the mean dissimilarity values suggests that the extinction-related change in CS− representation is established early and does not change over time, whereas the CS+ representation is initially more similar to its representation during Late Acquisition and then changes as extinction learning progresses. This interpretation is supported statistically by a repeated measures ANOVA, which revealed a significant CS Type by Time interaction (CS Type * Time $F_{(1,42)}$ = 4.8, p = 0.03; Table 1). As predicted, post-hoc *t*-tests averaging across ROIs (as there was no significant ROI x Time term in the ANOVA results) indicated this interaction was driven by a significant increase in CS+ dissimilarity from Early to Late Extinction (t_{42} = −4.6, p < 0.001) and a lack of significant change in CS− dissimilarity over the same time period (t_{42} = −1.1, p = 0.27).

Extinction-induced change in the multivariate representation of the CS+ in bilateral amygdala is moderated by individual differences in state anxiety

We reasoned that the extinction-induced change in representation of the CS+ would be greatest for individuals who exhibited the strongest conditioned fear response following acquisition training. However, we found no relationship between Late Acquisition SCRs and EL_{rsa+} in our *a priori* ROIs. We also expected that individuals with high state anxiety immediately following conditioning (see Figure S5 for all average reported anxiety values) would exhibit less change in the CS+ representation of the ROIs across time in Extinction. Of these regressions, one showed a significant correlation. EL_{rsa+} in bilateral amygdala was significantly (r^2 = 0.20, p = 0.015, Holm-corrected for n = 5 ROIs; Fig. 3) negatively correlated with reported state anxiety following Fear Acquisition. This result indicates that participants with higher state anxiety following conditioning exhibited a

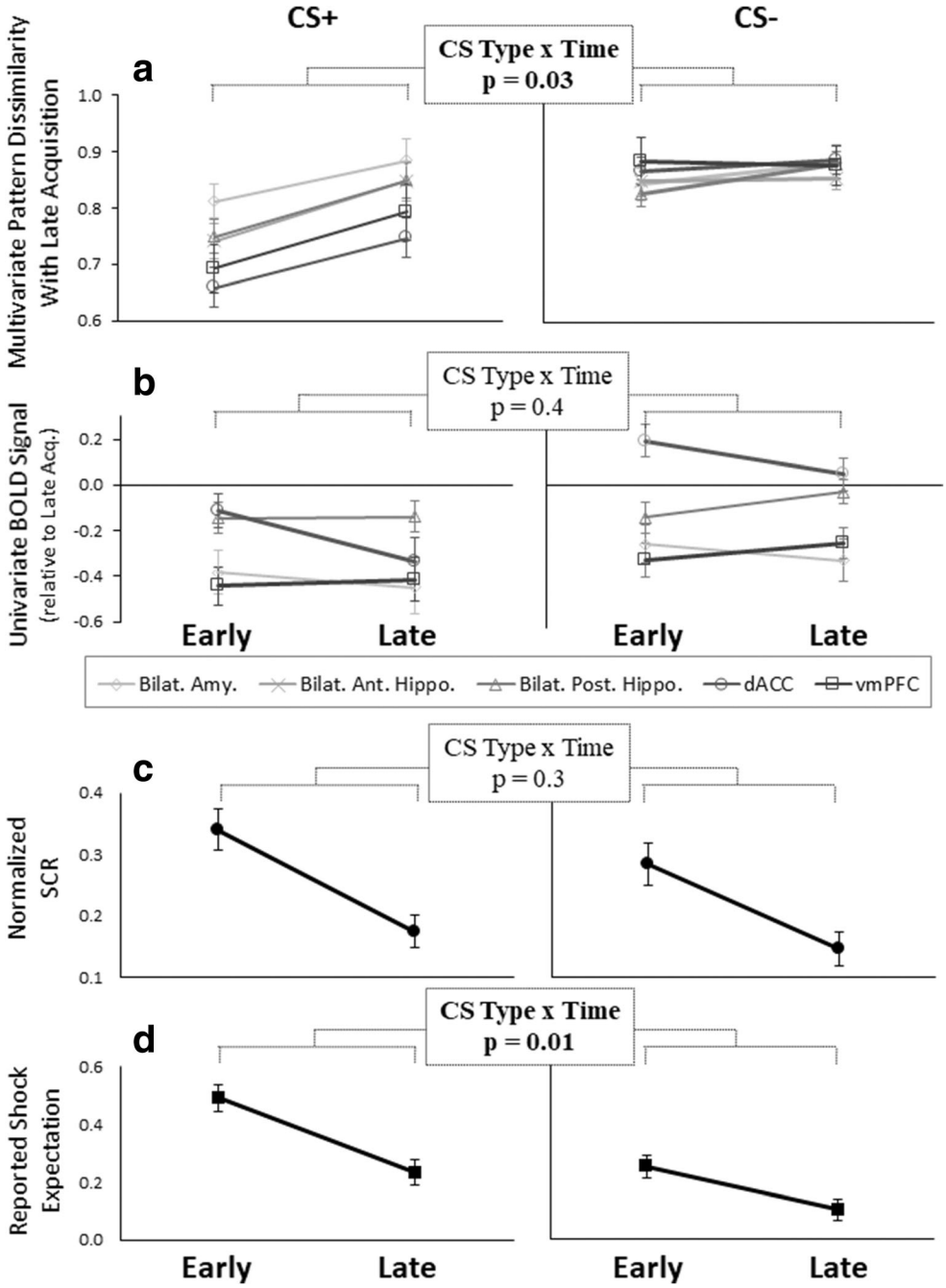

Fig. 2 Values of neuroimaging, psychophysiological, and behavioral metrics in Early and Late Extinction for CS+ and CS− stimuli. We highlight differences across measures in exhibiting the hypothesized CS Type by Time interactions. (A) The multivariate representational dissimilarity metric showed a significant CS Type by Time interaction. There was a greater difference between Early and Late Extinction for the reinforced conditioned stimulus (CS+; left plot) than for the nonreinforced stimulus (CS−; right plot), with the CS+ showing greater dissimilarity from Late Acquisition as extinction progressed. Values represent region of interest (ROI) means ± SEM. Bilat. Amy., Bilateral Amygdala; Bilat. Ant. Hippo., Bilateral Anterior Hippocampus; Bilat. Post. Hippo., Bilateral Posterior Hippocampus; dACC, dorsal Anterior Cingulate Cortex; vmPFC, ventromedial Prefrontal Cortex. (B) Average univariate BOLD response in the ROIs did not show a significant CS Type by Time interaction, although a main effect of region was found. (C) Normalized skin conductance response (SCR) also did not show a significant CS Type by Time interaction through Extinction, although main effects of CS Type and Time were found. (D) Reported shock expectancy showed a significant CS Type by Time interaction through Extinction; participants' expectation of shock to CS+ stimuli began higher and dropped off more significantly than did their expectation of shock to CS− stimuli

more persistent representation of the CS+ in the bilateral amygdala. None of the regressions of reported anxiety, SCR metrics, or shock expectancy with EL_{rsa}- showed a significant relationship (see Table S2).

Table 1 Three-factor ANOVA of multivariate dissimilarity

Model term	Degrees of freedom	F	Partial Eta^2	*p* value
Region	4	2.4	0.20	0.07
CSType	1	7.6	0.15	0.008
Time	1	13.7	0.25	0.001
Region*CSType	4	3.7	0.27	0.013
Region*Time	4	0.52	0.05	0.72
CSType*Time	1	4.8	0.10	0.034
Region*CSType*Time	4	1.4	0.12	0.26

Statistics for the three-factor ANOVA of pattern dissimilarity metrics. Factors are CS Type (CS+/CS−), Time in Extinction (Early/Late), and Brain Region (bilateral amygdala, bilateral posterior hippocampus, bilateral anterior hippocampus, dorsal anterior cingulate cortex, and ventral medial prefrontal cortex)

Multivariate dissimilarity changes through extinction show sensitivity to stimulus type when SCR measures do not, but do not track shock expectancy

As reported above, the multivariate dissimilarity metrics behaved differently for the CS+ and CS− events through extinction, with the CS+ dissimilarity values starting lower and becoming more dissimilar over time than those of the CS− events (Fig. 2A; Table 1). The 2-factor repeated measures ANOVA of the SCR data in extinction showed significant main effects of CS type ($F_{(1,42)}$ = 6.0, p = 0.018) and time ($F_{(1,42)}$ = 29.0, p < 0.001; Table 2), but no significant CS type-by-time interaction (CSType * Time $F_{(1,42)}$ = 1.4, p = 0.25). Although SCR magnitude decreased over the course of extinction, this decrease was present for both CS+ and CS− responses, with no differential effects (Early > Late paired t_{42} = 5.4, p < 0.001; Fig. 2C). The shock expectancy data also showed a significant CS Type-by-Time interaction (CS Type * Time $F_{(1,42)}$ = 6.4, p = 0.015; Table 3), with a greater reduction in ratings to the CS+ than the CS− as extinction progressed (Fig. 2D).

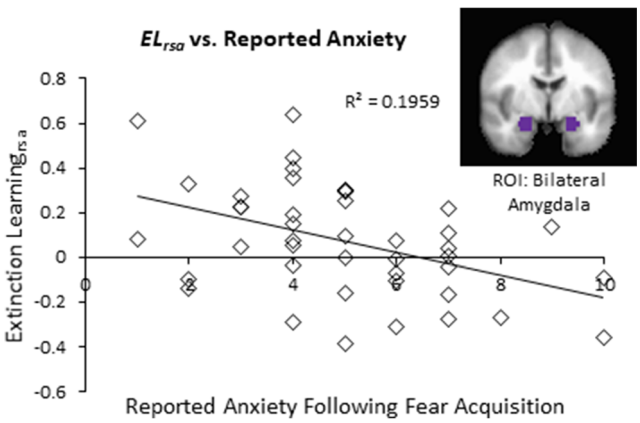

Fig. 3 Self-reported state anxiety immediately following fear acquisition was significantly negatively correlated (r^2 = 0.20, p = 0.015) with the extinction-related change in CS+ representational similarity in the amygdala. Higher anxiety across participants was associated with lower multivariate evidence of extinction learning (EL_{rsa})

To investigate the relationship between EL_{rsa+} and SCR and shock expectancy across individuals, regression models were created (one per ROI) to attempt to predict EL_{rsa+} from the change in these other variables between Early and Late Extinction. EL_{rsa+} was regressed against (SCR_{late} - SCR_{early}), where SCR_{late} and SCR_{early} each refer to the average normalized SCR ((SCR_{CS+} - SCR_{CS-})/MAX_{US})) across the associated time period of extinction, and EL_{rsa+} was regressed against SE_{late} - SE_{early}, where SE_{late} and SE_{early} each refer to the average shock expectancy to the CS+ across the associated time period in extinction. None of the regressions showed a significant relationship after correction for multiple comparisons (see Table S2 for uncorrected *p*-values).

Whole-brain searchlight RSA reveals additional brain regions with extinction-induced representation changes beyond the core extinction circuit model

The whole-brain, voxel-wise searchlight RSA revealed several additional brain regions where the neural representation of CS+ stimuli, relative to Late Acquisition, changed significantly between Early and Late Extinction. These regions included the mid-cingulate cortex, bilateral insular cortex/frontal operculum, right somatosensory cortex, left cerebellum, and multiple sectors of visual cortex. The sizes of these clusters, their corrected *p*-values (alphas), and the locations of their centers within the brain (as determined by AFNI's *whereami*

Table 2 SCR Two-factor ANOVA

Model term	Degrees of freedom	F	Partial Eta^2	*p* value
CSType	1	6.0	0.13	0.018
Time	1	29.0	0.41	<0.001
CSType*Time	1	1.4	0.03	0.25

Statistics for the two-factor ANOVA of normalized SCR (SCRs for CS+ and CS- divided by the maximum US SCR for each participant) in Extinction. Factors are CS Type (CS+/CS−) and Time in Extinction (Early/Late)

Table 3 Shock expectancy two-factor ANOVA

Model term	Degrees of freedom (*df*)	$F_{(df,42)}$	*p* value
CSType	1	18	<0.001
Time	1	75.2	<0.001
CSType*Time	1	6.4	0.015

Statistics for the two-factor ANOVA of shock expectancy data. Factors are CS Type (CS+/CS−) and Time in Extinction (Early/Late)

function) are included in Table 4. Some of these clusters are depicted in Fig. 4. The searchlight analysis of the CS− data found no regions in which the neural representation of CS− stimuli, relative to Late Acquisition, changed significantly between Early and Late Extinction, after correction for multiple comparisons.

Multivariate analysis yields additional insights relative to standard univariate fMRI extinction metrics

The repeated measures ANOVA on the univariate metrics had three significant terms: the main effect of ROI ($F_{(3,42)} = 9.1$, $p < 0.001$), the main effect of CS-type ($F_{(1,42)}=7.1$, $p = 0.01$), and the ROI-by-time interaction term ($F_{(3,42)} = 11.2$, $p < 0.001$; Table 5). Notably, the CS-type-by-time interaction term was not significant, suggesting changes in the BOLD signal between Early and Late Extinction are the same for the CS+ and CS−. Regression of the univariate equivalent of EL_{rsa+} (average CS+ activation in Late Extinction minus the average CS+ activation in Early Extinction) for each ROI against Late Acquisition SCR and reported anxiety following Late Acquisition showed no significant correlations. Therefore, the multivariate metrics reveal hypothesized CS-type-by-time interactions and sensitivity to individual differences in state anxiety that were not found using traditional univariate analytic approaches.

Discussion

The goal of this study was to characterize how extinction learning changes the representation of a fear-conditioned stimulus in the core neural circuit implicated in animal conditioning models. To achieve this goal, we derived a multivariate metric that first extracted a stable estimate of the representation of the CS+ from the Late Acquisition training phase, and then investigated how it changed across early and late phases of extinction training. A similar process was conducted for the explicitly-unreinforced stimulus (CS−). Extinction followed acquisition training in the same fMRI testing session but was conducted in a different 3D virtual environment.

We found that the contextual switch from Late Acquisition to Early Extinction training yielded a change in the neural representation of the CS- within all ROIs, spanning the amygdala, hippocampus, vmPFC and dACC. However, the representation of the CS+ resisted this change, becoming more dissimilar from Late Acquisition as extinction learning progressed to its later stage in the same ROIs. An analogous analysis using the univariate signal from our ROIs showed no CS type X time interaction during extinction, indicating that the multivariate analysis is more sensitive to such extinction-induced shifts in CS+ signaling. A whole-brain searchlight analysis yielded similar findings for the CS+ in some brain areas outside of this core extinction network, including the mid-cingulate gyrus, insula/frontal operculum, cerebellum, somatosensory cortex, and several sectors of the visual cortex. Providing evidence that this persistence in the CS+ representation early in extinction was related to its acquired threat value, we found that individuals who reported higher state anxiety at the end of acquisition training exhibited less change in their multivariate amygdala patterning of the CS+ during extinction. This relationship with state anxiety also was not evident in the univariate analysis.

As discussed in the Introduction, prior univariate fMRI studies of extinction in healthy adults have yielded conflicting

Table 4 Whole-brain searchlight cluster results

Center of mass (mm, MNI-152)						
X	Y	Z	Volume (ml)	Cluster α	Max Z-stat	CA_ML_18_MNIA: Macro Labels
46	−4	8	1.5	<0.01	4.5	Right Rolandic operculum (right insula)
−30	−66	−38	1.1	<0.01	4.6	Left cerebellum (Crus 2)
0	−18	42	0.86	<0.02	5.4	Mid-cingulate cortex
0	−82	−4	0.77	<0.03	4.2	Calcarine gyrus
−52	8	2	0.73	<0.03	4.3	Left Rolandic operculum
−14	−46	−8	0.65	<0.04	4.2	Left lingual gyrus
62	−10	34	0.60	<0.05	4.2	Right postcentral gyrus

Whole-brain voxel-wise searchlight results from the representational similarity analysis reflecting extinction learning (EL_{rsa}). The neural representation of acquired fear to the CS+ stimuli in these regions significantly changes from early to late extinction. Reported locations are for the center-of-mass of each cluster. Reported cluster multiple-comparison-corrected alphas are based on a set voxel-wise *t*-test *p*-value of 0.001

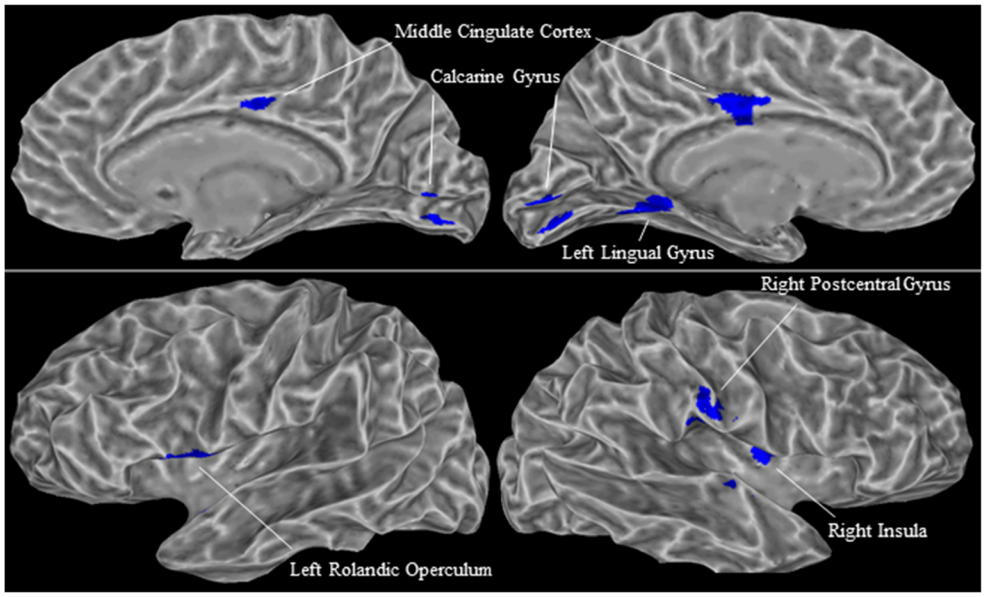

Fig. 4 Whole-brain voxel-wise searchlight results. Shown in blue are clusters of voxels in which EL_{rsa} was significant. These areas indicate where the multivariate neural representation of CS+ stimuli, relative to Late Acquisition, changed significantly between Early and Late Extinction. An additional region in the cerebellum is not shown

findings. Fullana et al.'s 2018 meta-analysis reported that univariate fMRI studies of within-session extinction learning do not yield consistent changes in the vmPFC-amygdala-hippocampal circuit. Instead, extinction learning reliably activated other frontolimbic and subcortical structures (e.g., dACC, insula, striatum, thalamus, midbrain), as well as visual and somatosensory regions when extinction training was accompanied by a context shift. Fullana et al. (2018) noted the similarity of these results to meta-analyses of fear acquisition (e.g., Fullana et al., 2016) and speculated that some of these regions may maintain a representation of the acquired fear value of the CS+ through extinction.

Our approach directly addresses this latter point and some other limitations of the prior literature. Indeed, we found evidence for a carryover effect of the CS+ threat representation from Late Acquisition to Early Extinction training in several brain regions implicated in conditioned fear learning and extinction, but these representations change by Late Extinction (albeit less so in the amygdala for individuals with high state anxiety). Thus, the temporal dynamics and representational content of activity in these regions are important to consider, and studies that combine data across all extinction trials in these regions may be conflating threat and safety signals. We note that other brain regions may instead encode error prediction signals or other facets of extinction learning that are not captured by the change in representational similarity investigated here. By comparing univariate and multivariate results within our ROIs, we demonstrate the added value of the latter approach.

Rodent studies have shown that "fear" and "extinction" cells can be intermixed within the basolateral amygdala, along with neurons that signal both representations as a potential index of fear persistence (Herry et al., 2008). In addition, the prelimbic cortex contains both inhibitory and excitatory pathways that are differentially gated through the ventral hippocampus and basolateral amygdala to reduce or enhance conditioned fear expression, respectively (Sotres-Boyen et al., 2012). One difference between our results and the existing rodent literature is that we find evidence for extinction-related changes in threat signaling in both the dACC and vmPFC. In contrast, the rodent literature tends to find more consistent differences between these structures, with the dACC (prelimbic cortex) relating more to fear acquisition/expression/recovery and the vmPFC (infralimbic cortex) more involved with fear extinction and safety signaling (reviewed in Milad & Quirk, 2012). Nonetheless, rodent optogenetic studies have shown that these two prefrontal structures can directly interact with one another

Table 5 Three-factor ANOVA of univariate BOLD signal

Model term	Degrees of freedom	F	Partial Eta^2	*p* value
Region	3	9.1	0.41	<0.001
CSType	1	7.1	0.15	0.01
Time	1	0.81	0.02	0.38
Region*CSType	3	2.4	0.15	0.09
Region*Time	3	11.2	0.46	<0.001
CSType*Time	1	0.64	0.02	0.43
Region*CSType*Time	3	0.68	0.05	0.57

Statistics for the three-factor ANOVA of univariate average BOLD (blood oxygenation level dependent) fMRI signal change. Factors are CS Type (CS+/CS−), Time in Extinction (Early/Late), and Brain Region (bilateral amygdala, bilateral posterior hippocampus, dorsal anterior cingulate cortex, and ventral medial prefrontal cortex)

(Ji & Neugebauer, 2012). To be effective, extinction learning must engage specific inhibitory neuronal subtypes within the infralimbic cortex and amygdala; otherwise, fear behaviors continue to be expressed (reviewed in Courtin et al., 2013). Likewise, optogenetic manipulation of specific pathways within the prelimbic cortex can recover fear memories (Courtin et al., 2014), which increases theta synchrony across the ventral hippocampus, basolateral amygdala, and mPFC (Stujenske et al., 2014). All of these considerations suggest that a more refined spatial approach to fMRI data analysis is warranted to facilitate cross-species comparisons regarding the functional roles of these structures.

Multivariate analytic approaches to fMRI that do not apply spatial smoothing to the raw images should be more sensitive than traditional mass univariate approaches to detect these subtle patterns of threat and safety value within a prescribed anatomical region (although subcortical regions may require denser sampling and/or special multiple comparisons correction for small volumes). As shown here, changes in such neural representations can be extracted even in the absence of measurable behavioral fear expression during later stages of extinction training. It is perhaps not surprising that our multivariate extinction learning metric didn't track concurrent SCR/shock expectancy ratings during the extinction session, as animal studies show that the plasticity in this core circuit functionally relates primarily to initiating consolidation processes for the extinction learning episode rather than within-session expression of extinction learning (Milad & Quirk, 2012). However, we note that SCR is a noisy dependent measure, especially late in extinction learning where the data are skewed towards zero, and our crude subdivision of extinction phases using a split-half approach (early vs. late trials) may have obscured more precise relationships.

As mentioned, the present report constituted an exploratory analysis of previously collected data. Some limitations arise from the fact that the original paradigm was not optimally designed for trial-wise multivariate pattern analysis (such as that performed by Visser et al., 2011), and thus we grouped data across trial blocks (see *Supplemental Materials* for details). Specifically, the time interval between consecutive trials of each type was not constant throughout the paradigm and sufficiently long to minimize the influence of intrinsic noise correlations in the trial-wise estimates (see *Methods*). Future study designs with optimally timed stimulus presentations and the collection of fear acquisition and extinction data in the same imaging session would allow expansion of the current work. On the other hand, we note that most fear conditioning studies use variable trials to avoid temporal confounds in the prediction of CS onsets, and so our approach, while providing limited temporal precision, may be more generalizable to standard conditioning task designs and avoids inadvertent temporal conditioning due to fixed training intervals.

Our paradigm contained several salient shifts between acquisition and extinction that could help to establish a contextual boundary between training phases and promote new learning. As with all conditioning procedures, the extinction phase omitted the reinforcer which yields error prediction signals and creates a safe context. In addition, the 3D VR technology created a strong environmental context shift from an indoor to an outdoor setting (or vice-versa, counterbalanced across participants), which also entailed a brief temporal pause in fMRI scanning between experimental phases. Future studies that compare results with and without an environmental context shift are needed to determine the extent to which our results depend on this feature of the experimental design. In particular, it is interesting that the CS− representation in Early Extinction is already quite dissimilar to that from Late Acquisition. We suspect that the degree of dissimilarity in CS- signaling would be reduced if both phases were conducted in the same environmental context, but this conjecture requires experimental validation. Given its role in event boundary detection (Baldassano et al., 2017), it may be surprising that the hippocampus shows a temporal delay in signaling a change of the CS+ representation upon the context shift. One potential reason for this pattern is that the hippocampal signaling of the threat value of the CS+ overshadows that of the background context shift when activity is time-locked to the cue. Electrophysiological studies in rodents have shown that the ventral hippocampus, which is the region most directly connected to the amygdala, plays a role in CS+ signaling (Maren & Holt, 2004); conditioned fear stress exposure enhances hippocampal long-term potentiation (Inoue et al., 2013); and optogenetic activation of a hippocampal engram recovers conditioned fear memories (Liu et al., 2012). Cued conditioned threats are learned more quickly, exhibit larger behavioral responses, and extinguish more slowly than background contextual conditioning when assayed separately within the same task (LaBar & LeDoux, 1996). Alternatively, hippocampal coding of the context shift in the presence of a CS+ might be too transient to pick up using our trial-averaging technique, and/or the context signaling may be more evident when the CS+ is not present (e.g., during the intertrial interval).

Future studies should also endeavor to link these representational changes with specific cognitive and computational processes. Of the extinction-induced changes in our ROIs, only the amygdala's representation of the CS+ tracked individual differences in post-acquisition state anxiety. Extinction-induced changes in multivariate patterning in other brain regions may be driven, in turn, by the amygdala itself or may be related to other cognitive processes implicated in computational models of conditioning, such as attentional allocation, value updating, or contextual binding (Armony et al., 1995; Schmajuk, 1997).

Finally, we acknowledge that there is a current debate regarding how well the neural circuitry involved in defensive conditioning processes maps onto feelings of fear and anxiety,

with some researchers calling for a reframing of conditioning processes as reflecting survival-based threat detection computations that are broader than those relating to the experience of fear *per se* (LeDoux, 2012; LeDoux & Pine, 2016; but see Fanselow & Pennington, 2016). For instance, using a threat exposure paradigm, Taschereau-Dumouchel, Kawato & Lau, (2019) showed that, while there was some overlap in fMRI signals that predicted both subjective anxiety and SCR reactivity to evolutionary-based threats, other regions showed selectivity to one of these measures and not the other. In this paper, we chose to maintain the traditional terminology that refers to defensive Pavlovian training paradigms as "fear conditioning" to link our study to the broader extant literature, while recognizing that fMRI responses elicited to conditioned stimuli may not necessarily map onto those that mediate the subjective experience of fear. Interestingly, our results inform this debate by showing that a dissociation between conditioning processes and the subjective experience of fear/anxiety may not be so clear-cut. In particular, we found that individual differences in the subjective experience of anxiety at the end of acquisition training predicted a persistence in CS+ patterning in the amygdala during Early Extinction training. As discussed above, this relationship was not found to physiological expression of conditioned learning as measured by SCR. These results suggest that fMRI signaling of conditioned stimuli may reflect some integration of subjective aspects of fear/anxiety. Nonetheless, we note that this influence was selective to the amygdala ROI and thus may be more separable in other components of fear-conditioning circuitry. Future studies that assay emotions more comprehensively during fear conditioning are warranted to further interrogate the relationship between subjective experience and neural representations of conditioned stimuli.

Open practices statement The voxel-wise whole-brain searchlight RSA statistical map for the CS+ events (shown in Fig. 4) will be made available upon publication via the NeuroVault website (neurovault.org). Participant-wise RSA dissimilarity metrics for each ROI, average univariate activation values for each ROI, reported shock expectancy, reported state anxiety levels, and skin conductance response values will be made available upon publication via the Open Science Framework website (osf.io). The experiment was not preregistered.

References

Ahs, F., Kragel, P. A., Zielinski, D. J., Brady, R., & LaBar, K. S. (2015). Medial prefrontal pathways for the contextual regulation of extinguished fear in humans. *Neuroimage, 122*, 262–271.

Armony, J. L., Servan-Schreiber, D., Cohen, J. D., & LeDoux, J. E. (1995). An anatomically constrained neural network model of fear conditioning. *Behavioral Neuroscience, 109*, 246–257.

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., Norman, K. A. (2017). Discovering Event Structure in Continuous Narrative Perception and Memory. *Neuron, 95*, 709–721.

Ben-Yakov, A., Rubinson, M., & Dudai, Y. (2014). Shifting gears in hippocampus: Temporal dissociation between familiarity and novelty signatures in a single event. *The Journal of Neuroscience, 34,* 12973–12981.

Bouton, M. E. (1993). Context, time, and memory retrieval in the interference paradigms of Pavlovian learning. *Psychological Bulletin, 114*, 80-99.

Bukalo, O., Pinard, C. R., Silverstein, S., Brehm, C., Hartley, N. D., Whittle, N., Colacicco, G., Busch, E., Patel, S., Singewald, N., & Holmes, A. (2015). Prefrontal inputs to the amygdala instruct fear extinction memory formation. *Science Advances, 1,* e1500251.

Courtin, J., Bienvenu, T. C. M., Einarsson, E. O., & Herry, C. (2013). Medial prefrontal cortex neuronal circuits in fear behavior. *Neuroscience, 240,* 219–242.

Courtin, J., Chaudun, F., Rozeske, R. R., Karalis, N., Gonzalez-Campo, C., Wurtz, H., et al. (2014). Prefrontal parvalbumin interneurons shape neuronal activity to drive fear expression. *Nature, 505,* 92–96.

Cox, R. W. (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research, 29*, 162–173.

Dejean, C., Courtin, J., Rozeske, R. R., Bonnet, M. C., Dousset, V., Michelet, T., & Herry, C. (2015). Neuronal circuits for fear expression and recovery: Recent advances and potential therapeutic strategies. *Biological Psychiatry, 78,* 298–306.

Do-Monte, F. H., Manzano-Nieves, G., Quinones-Laracuente, K., Ramos-Medina, L., & Quirk, G. J. (2015). Revisiting the role of infralimbic cortex in fear extinction with optogenetics. *Journal of Neuroscience, 35,* 3607–3615.

Dunsmoor, J. E., Niv, Y., Daw, N., & Phelps, E. A. (2015). Rethinking extinction. *Neuron, 88,* 47–63.

Fanselow, M. S. & Pennington, Z. T. (2016). The danger of LeDoux and Pine's two-system framework for fear. *American Journal of Psychiatry, 173,* 1120–1121.

Fullana, M. A., Albajes-Eizagirre, A., Soriano-Mas, C., Vervliet, B., Cardoner, N., Benet, O., Radua, J., & Harrison, B. J. (2018). Fear extinction in the human brain: a meta-analysis of fMRI studies in healthy participants. *Neuroscience and Biobehavioral Reviews, 88,* 16–25.

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

Green, S. R., Kragel, P. A., Fecteau, M. E., & LaBar, K. S. (2014). Development and validation of an unsupervised scoring system (Autonomate) for skin conductance response analysis. *International Journal of Psychophysiology, 91*, 186–193.

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.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics, 6*, 65–70.

Inoue, S., Kamiyama, H., Matsumoto, M., Yanagawa, Y., Hiraide, S., Saito, Y., Shimamura, K., & Togashi, H. (2013). Synaptic modulation via basolateral amygdala on the rat hippocampus-medial prefrontal cortex pathway in fear extinction. *Journal of Pharmacological Sciences, 123,* 267–278.

Ji, G., & Neugebauer, V. (2012). Modulation of medial prefrontal cortical activity using in vivo recordings and optogenetics. *Molecular Brain, 5,* 36.

Ji, J. & Maren, S. (2007). Hippocampal involvement in contextual modulation of fear extinction. *Hippocampus, 17,* 749–758.

Kriegeskorte, N. (2011). Pattern-information analysis: From stimulus decoding to computational-model testing. *NeuroImage 56*(2), 411–421.

LaBar, K. S., Gatenby, J. C., Gore, J. C., LeDoux, J. E., & Phelps, E. A. (1998). Human amygdala activation during conditioned fear acquisition and extinction: A mixed-trial fMRI study. *Neuron*, *20,* 937–945.

LaBar, K. S., & LeDoux, J. E. (1996). Partial disruption of fear conditioning in rats with unilateral amygdala damage: Correspondence with unilateral temporal lobectomy in humans. *Behavioral Neuroscience, 110,* 991–997.

Laurent, V., & Westbrook, R. F. (2009). Inactivation of the infralimbic but not the prelimbic cortex impairs consolidation and retrieval of fear extinction. *Learning & Memory, 16*, 520–529.

LeDoux, J. E. (2012). Rethinking the emotional brain. *Neuron, 73,* 653–676.

LeDoux, J. E. & Pine, D. S. (2016). Using neuroscience to help understand fear and anxiety: a two-system framework. *American Journal of Psychiatry, 173,* 1083–1093.

Lissek, S. (2012). Toward an account of clinical anxiety predicated on basic, neurally mapped mechanisms of Pavlovian fear-learning: the case for conditioned overgeneralization. *Depression and Anxiety, 29*, 257–263.

Liu, X., Ramirez, S., Pang, P. T., Puryear, C. B., Govindarajan, A., Deisseroth, K., & Tonegawa, S. (2012). Optogenetic stimulation of a hippocampal engram activates fear memory recall. *Nature, 484,* 381–385.

Maldjian, J. A., Laurienti, P. J., Kraft, R. A., & Burdette, J. H. (2003). An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *Neuroimage, 19*, 1233–1239.

Maren, S. (2011). Seeking a Spotless Mind: Extinction, Deconsolidation, and Erasure of Fear Memory. *Neuron, 70*, 830–845.

Maren, S. & Holt, W. G. (2004). Hippocampus and Pavlovian fear conditioning in rats: Muscimol infusions into the ventral, but not dorsal, hippocampus impair the acquisition of conditional freezing to an auditory conditional stimulus. *Behavioral Neuroscience, 118,* 97–110.

Marks, I., & Tobena, A. (1990). Learning and unlearning fear: a clinical and evolutionary perspective. *Neuroscience and Biobehavioral Reviews, 14*, 365–384.

Milad, M. R., Quirk, G. J. (2012). Fear extinction as a model for translational neuroscience: ten years of progress. *Annual Review of Psychology*, *63*, 129–151.

Milad, M. R., Wright, C. I., Orr, S. P., Pitman, R. K., Quirk, G. J., & Rauch, S. L. (2007). Recall of fear extinction in humans activates the ventromedial prefrontal cortex and hippocampus in concert. *Biological Psychiatry, 62,* 446–454.

Phelps, E. A., Delgado, M. R., Nearing, K. I., & LeDoux, J. E. (2004). Extinction learning in humans: role of the amygdala and vmPFC. *Neuron, 43,* 897–905.

Poppenk J, Evensmoen HR, Moscovitch M, Nadel L. (2013). Long-axis specialization of the human hippocampus. *Trends in Cognitive Sciences*, *17*, 230–240

Radvansky, G. A. & Zacks, J. M. (2017). Event boundaries in memory and cognition. *Current Opinion in Behavioral Sciences, 17,* 133–140.

Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., & Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences USA, 98*, 676–682.

Rozeske, R. R., & Herry, C. (2018). Neuronal coding mechanisms mediating fear behavior. *Current Opinion in Neurobiology, 52*, 60–64.

Schmajuk, N. A. (1997). *Animal learning and cognition: a neural network approach.* New York: Cambridge University Press.

Sehlmeyer, C., Dannlowski, U., Schoning, S., Kugel, H., Pyka, M., Pfleiderer, B., Zwitserlood, P., Schiffbauer, H., Heindel, W., Arolt, V., Konrad, C. (2011). Neural Correlates of Trait Anxiety in Fear Extinction. *Psychological Medicine, 41(4)*, 789–98.

Sevenster, D., Visser, R. M., & D'Hooge, R. (2018). A translational perspective on neural circuits of fear extinction: Current promises and challenges. *Neurobiology of Learning and Memory, 155,* 113–126.

Sierra-Mercardo, D., Padilla-Coreano, N., & Quirk, G. J. (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.

Sotres-Boyen, F., Sierra-Mercado, D., Pardilla-Delgado, E., & Quirk, G. J. (2012). Gating of fear in prelimbic cortex by hippocampal and amygdala inputs. *Neuron, 76(4),* 804–812.

Stujenske, J. M., Likhtik, E., Topiwala, M. A., & Gordon, J. A. (2014). Fear and safety engage competing patterns of theta-gamma coupling in the basolateral amygdala. *Neuron, 83,* 919–933.

Taschereau-Dumouchel, V., Kawato, M., & Lau, H. (2019). Multivoxel pattern analysis reveals dissociations between subjective fear and its physiological correlates. Molecular Psychiatry, *in press*.

Tong, F. & Pratte, M. S. (2012). Decoding patterns of human brain activity. *Annual Review of Psychology, 63,* 483–509.

Visser, R. M., Haan, M. I. C., Beemsterboer, T., Haver, P., Kindt, M., & Scholte, H. S. (2016). Quantifying learning-dependent changes in the brain: Single-trial multivoxel pattern analysis requires slow event-related fMRI. *Psychophysiology, 53*, 1117–1127.

Visser, R. M., Kunze, A. E., Westhoff, B., Scholte, H. S., & Kindt, M. (2015). Representational similarity analysis offers a preview of the noradrenergic modulation of long-term fear memory at the time of encoding. *Psychoneuroendocrinology, 55,* 8–20

Visser, R. M., Scholte, H.S., Beemsterboer, T., & Kindt, M. (2013). Neural pattern similarity predicts long-term fear memory. *Nature Neuroscience, 16(4)*, 388–390.

Visser, R. M., Scholte, H. S., & Kindt, M. (2011). Associative Learning Increases Trial-by-Trial Similarity of BOLD-MRI Patterns. *The Journal of Neuroscience, 31(33)*, 12021–12028.

Xu, J. (2015) Implications of cortical balanced excitation and inhibition, functional heterogeneity, and sparseness of neuronal activity in fMRI. *Neuroscience and Biobehavioral Reviews, 57,* 264–270.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: a mind-brain perspective. *Psychological Bulletin, 133,* 273–293.