Distinct representational properties of cues and contexts shape fear learning and extinction

Antoine Bouyeure[1], Daniel Pacheco[1], Marie-Christin Fellner[1], George Jacob[1], Malte Kobelt[1], Jonas Rose[2], Nikolai Axmacher[1]

1. Department of Neuropsychology, Ruhr-Universität Bochum, Bochum 44801, North Rhine-Westphalia, Germany
2. Department of Neuroscience, Ruhr-Universität Bochum, Bochum 44801, North Rhine-Westphalia, Germany

Correspondance to: antoine.bouyeure@rub.de ; nikolai.axmacher@rub.de

Abstract

Extinction learning does not erase previously established memories but inhibits the expression of fear by the formation of new memory traces that are strongly context-dependent. Previous human neuroimaging studies using representational similarity analysis revealed several core properties of memory traces during fear learning, including their tendency to generalize beyond the initial context – a process described as "cue generalization" – and their reliance on sensory rather than conceptual representational formats. How fear memories are altered during extinction learning, however, remains largely unknown. To address this question, we used a novel experimental paradigm involving multiple cues and contexts in each experimental phase, which allowed us to disentangle the effect of contingency changes (i.e., reversal learning) from the disappearance of unconditioned stimuli during extinction learning. Our data show that contingency changes during reversal induce memory traces with distinct representational geometries characterized by stable activity patterns across repetitions in the precuneus, which interact with specific context representations in medial and lateral prefrontal cortex. The representational geometries of these traces differ strikingly from the generalized patterns established during initial fear learning and persist in the absence of an unconditioned stimulus during extinction. Interestingly, increased levels of prefrontal context specificity predict the subsequent reinstatement of fear memory traces, providing a possible mechanistic explanation for the clinical phenomenon of fear renewal. Our findings show that contingency changes induce novel memory traces with distinct representational properties that are reminiscent to those observed during episodic memory formation and contrast with the generalized representations of initial fear memories. These results shed new light on the neural mechanisms underlying the malleability of memories that support cognitive flexibility, and contribute to conceptual frameworks of extinction learning during the treatment of anxiety disorders.

1. Introduction

Fear acquisition refers to the process of learning the association between a neutral conditioned stimulus (CS) and an aversive unconditioned stimulus (US). It is typically a rapid and robust process that may create long-lasting fear memories which may persist after the threats have passed. This persistence can be evolutionarily advantageous, since it may be adaptive to rather respond to a false alarm than to miss a potential threat. However, the inability to eventually suppress a fear response in the absence of actual danger can become dysfunctional and has been proposed as a key etiological factor in conditions such as anxiety disorders and post-traumatic stress disorder (Milad & Quirk, 2012).

While fear acquisition is rapid and robust, the suppression of fear responses in the absence of the US – i.e., fear extinction – is strongly context-dependent and more flexible (Maren et al., 2013; Milad & Quirk, 2012; Liu et al., 2024). This is demonstrated by the phenomena of spontaneous recovery, renewal, and reinstatement, all of which show that the original fear memory can resurface under certain conditions (Bouton, 2002). Specifically, fear renewal reflects a return of fear responses after a change in context, showing that extinction does not erase the original fear memory trace but inhibits it selectively within the extinction context (Greco & Liberzon, 2016). Learning and extinction do not occur solely in relation to fear, but also during processes of reinforcement learning and reversal, i.e., related to contingency changes more generally (Schiller et al., 2008; Wisniewski et al., 2023). In these cases, the context dependency of extinction may support cognitive flexibility since appropriate actions can be selected according to situational demands (Schiller & Delgado, 2010; Chaby et al., 2019; Xin et al., 2024). Contrastingly, the context specificity of extinction learning may be detrimental during treatments of anxiety disorders if therapy-induced fear reductions do not generalize beyond the therapeutic setting (Maren et al., 2013).

Much of our fundamental understanding of the formation of fear memory traces and their suppression during extinction learning has been derived from optogenetic studies in rodents, which describe the formation and modification of fear engrams with valence and context representations in the amygdala and hippocampus, respectively (Liu et al., 2015; Josselyn et al., 2015; Redondo et al., 2014). These studies further showed that extinction learning depends on plasticity of hippocampal context representations (Redondo et al., 2014) as well as on prefrontal cortex engrams (Ramanathan et al., 2018; Gu et al., 2022; Lissek & Tegenthoff, 2024).

In humans, neuroimaging studies have reported activation of a canonical “fear network” during acquisition (with prominent roles of the dorsal anterior cingulate cortex and insula, and a more inconsistent role of the amygdala; Fullana et al., 2016) and recruitment of the hippocampus and ventromedial prefrontal cortex during extinction (Fullana et al., 2016; Maren et al., 2013), putatively reflecting context dependency and safety learning, respectively (Maren, Phan, & Liberzon, 2013). Indeed, meta-analyses have shown that despite some moderate overlap, the brain regions involved in extinction learning differ substantially from those involved in fear acquisition (Maren et al., 2013). Moreover, reversal – involving a change in contingencies rather than the mere absence of a US – particularly engages regions involved in prediction error detection and cognitive flexibility, such as the dorsomedial and lateral prefrontal cortex (Wisniewski et al., 2023; Xin et al., 2024). This points towards the inhibition of previously threatening stimuli via executive control during reversal, a process not typically observed in standard extinction paradigms.

While functional magnetic resonance imaging (fMRI) activation studies have been instrumental in identifying the brain regions involved in fear learning and extinction, they are insensitive to the patterns of neural activity that underlie the stimulus-specific representations of threat cues and contexts. By contrast, representational similarity analysis (RSA; Kriegeskorte et al., 2008) has emerged as a powerful tool to track the fate of distinct memory traces over time (Rissman and Wagner, 2012; Heinen et al., 2024). This method has provided important novel insights into the representational signatures and “geometries” (i.e., patterns of representational distances among items) that support the formation, stabilization, and possible subsequent refinement and modification of memory traces. This has informed our understanding of the basic mechanisms of learning and memory, while also contributing to more mechanistic theories of memory distortions in mental disorders. For example, Visser et al. (2011, 2013) demonstrated that trial-by-trial similarities of blood oxygen level-dependent (BOLD) patterns increase during associative learning in regions of the fear network such as the anterior cingulate cortex (ACC), ventromedial prefrontal cortex (vmPFC), or superior frontal cortex. Similar representational signatures of “cue generalization” – i.e., increasing levels of similarity among the memory traces of different items associated with the same valence – were observed in the amygdala related to memories of a stressful episode (Bierbrauer et al., 2021), as well as in sensory regions and areas of the salience network for aversive trauma-analogue stimuli (Kobelt et al., 2024). Further, RSA can be used to study how specific neural patterns are reactivated during memory, a mechanism also referred to as “encoding-retrieval similarity” (e.g., Kobelt et al., 2024). For example, Hennings et al. (2022) showed a selective reactivation of fear versus extinction memories in the medial PFC and hippocampus depending on encoding context. Furthermore, the similarity between the neural patterns that are elicited across different presentations of a given item (within-item similarity) describes how stable the representation of a particular item is, regardless of its valence, and has been linked to episodic memory performance (Xue et al., 2010). Finally, the difference between within-item stability and between-item generalization, commonly referred to as “specificity” (Xue et al., 2010; Xue et al., 2013; Zheng et al., 2018; Sommer et al., 2022), quantifies the amount of item-specific information in a representation. This representational property could be particularly fruitful as a means to study the influence of fear reversal or extinction on context representations, which have never been analyzed in previous fear and extinction learning studies.

Here, we aimed to systematically investigate how the neural representations of cues and contexts change across different phases of learning, including acquisition, reversal, and two consecutive test phases with new contexts and previous contexts, respectively, in which all cues are extinguished. We presented the CS cues in each phase in multiple different contexts that changed between phases, which allowed us to study the role of context specificity by comparing the similarity between same vs. different contexts in each phase (see Figure 1).

We hypothesized that the representational geometry of CS cues changes across learning phases, reflecting the inhibition of fear memories during reversal, as well as the formation of novel memories of cues with updated contingencies. More specifically, we expected cue generalization effects in regions of the fear network, item stability in areas related to episodic memory, and context-specific representations in the hippocampus and PFC. Finally, we hypothesized context specificity during reversal to influence the reinstatement of fear memory traces during the test phases.

2. Methods

Participants

Thirty healthy participants were recruited via flyers and the online recruitment systems of the Faculty of Psychology at Ruhr University Bochum. As our paradigm consisted of a two-day design with two experimental phases per day, we observed some attrition between experimental phases. The number of participants with usable fMRI data for each phase was as follows: N = 30 for the first phase of day one, N = 29 for the second phase of day one, N = 27 for the first phase of day two, and N = 26 for the second phase of day two.

fMRI data were considered unusable in case of incomplete data (i.e., absence of data for all four phases, n = 9) or significant head movement (>2.5 mm in any direction, n = 2). The final sample for fMRI analysis included 24 participants (8 males) between 18 and 32 years of age (mean: 24.69 years, standard deviation: 3.6). All participants provided written informed consent before participation and were unaware of the aims of the experiment. The procedures were performed in accordance with the tenets of the Declaration of

Helsinki and were approved by the ethical review board of the Faculty of Psychology at Ruhr University Bochum.

General procedure and stimuli

The paradigm was administered to participants in the MRI scanner and consisted of four experimental phases spanning two days. To make the experiment more engaging for participants, they were presented with the narrative of “Nina the unlucky backpaper” and asked to play as the character Nina. During her fictitious trip, Nina would visit different places represented by videos of natural scenes that served as contexts. In each of these places, Nina would interact with different household appliances (the CS). Due to her misfortune, many of these items contain a serious defect and their manipulation could result in a mild electric shock (the US) experienced by Nina, and by extension, the participant. The four experimental phases therefore correspond to different trips undertaken by Nina during her travels.

Each phase comprised 128 trials with a similar structure: presentation of a context (video showing a natural scene) for 2 seconds, followed by the CS (household appliance) embedded within the context for 1 second. US expectancy responses were then collected during a 2.5-s periodusing a 4-point Likert scale, followed by the delivery (or absence) of an electric shock (US). Finally, participants saw a fixation cross which served as an inter-stimulus interval (Figure 1A). A total of eight CSs were presented during each phase and the same CSs were shown in all phases. The experimental phases differed in the way the CSs were associated with a US as well as in the possible contexts in which the CSs were embedded.

Visual stimuli were presented using the Presentation software package (Neurobehavioral Systems, Berkeley, CA, USA). Electrical stimulation was delivered using a constant voltage stimulator (STM2000, BIOPAC Systems, Goleta, CA, USA) with electrodes attached to the index finger. The intensity of the electrical stimulation was adjusted individually for each participant prior to the fear acquisition phase until the participants rated the sensation as unpleasant but not painful. A fixation cross was displayed with a jittered duration (7–9 s) to serve as an intertrial interval at the end of each trial. Each experimental phase contained 128 trials in total.

Fear acquisition, reversal, and test

The first day of the experiment comprised two phases: fear acquisition and fear reversal (Figure 1B). During the fear acquisition phase, four CSs were associated with a US (CS+, each with 50% reinforcement rate), while the other four were never followed by a US (CS-). Every CS was associated with four different contexts (natural scenes) with equal likelihood (i.e., across the 128 trials of each phase, every combination of a given CS with a given context occurred four times).

The fear acquisition phase was immediately followed by the fear reversal phase. Here, participants were again presented with the same CS. However, half of the CSs that were associated with a US during fear acquisition were no longer associated with a US (CS+-), while the other half remained associated with a US (CS++). Similarly, half of the CS- cues became associated with a US (CS-+), while the other half remained unassociated with a US (CS--). Both CS++ and CS-+ were reinforced in 50% of the trials.

The second day of the experiment comprised two experimental test phases: test in new contexts and test in acquisition/reversal contexts (Figure 1B). During these two test phases, no CS was ever associated with a US. They differed in terms of the context videos in which the CS were embedded, with new contexts for the “test_{new}” phase, and the previous acquisition and reversal contexts for the “test_{old}” phase (see details below).

To summarize: CS++ cues were associated with a US during both fear acquisition and reversal; CS-+ cues were not associated with a US during fear acquisition but were during reversal; CS+- cues were associated with a US during fear acquisition but not during reversal; CS-- cues were not associated with a US during either fear acquisition or reversal.

Relationship between CS types, contexts, and experimental phases

In all phases, the context videos and CS types were presented in different pseudorandom orders, such that they were orthogonal to each other. The first and last trials of each participant contained unreinforced cues. The assignment of each cue to the four possible CS types was counterbalanced across participants, as was the assignment of the type of videos used as context during each experimental phase.

Regarding context: One set of four videos was shown during fear acquisition (set “A”); a new set of four videos was shown during fear reversal (set “B”); a third set of eight videos was shown during test in new contexts (set “C”); during test in previous contexts, the four videos shown during fear acquisition and the four videos shown during fear reversal were shown (sets “A” and “B”, in pseudorandom order).

Thus, the two test phases differed only in the type of context videos shown (new contexts during test$_{\text{new}}$ and previously shown contexts test$_{\text{old}}$).

Behavioral data analysis

Participants provided US expectancy ratings after the presentation of each CS by indicating on a 4-point Likert scale how dangerous or safe they perceived the CS to be. We examined the influence of experimental phase and CS type on US expectancy by averaging US expectancy ratings across all trials of a given CS type, separately for each experimental phase and participant. This resulted in four (averaged) US expectancy ratings per experimental phase per participant.

As the repeated measures ANOVA assumption of sphericity was not met by the data (Mauchly test; W=0.39, p<0.0001), we chose a linear mixed effects (LME) approach to predict the effect of cue type and learning phase on US expectancy, with subject as a random effect. We used Satterthwaite's approximation of degrees of freedom, as implemented in the R packages lme4 and lmerTest, using the restricted maximum likelihood method. Studies have shown that this provides a more robust estimate of degrees of freedom, and thus reduces the risk of type 1 error, compared to other methods such as the likelihood ratio test (Luke, 2017).

As a preview of our behavioral data analyses, the results showed that participants quickly learned the contingencies in the initial fear acquisition phase, as well as the contingency changes introduced between the subsequent reversal and test$_{new}$/test$_{old}$ phases (Figure 1B).

MRI data collection

All MRI data were acquired on a Philips 3T Achieva scanner 3T MRI scanner (Philips Healthcare, Best, Netherlands). MRI data were acquired with simultaneous recording of electroencephalographic data and skin conductance recording in the scanner, which are not presented here. A reference structural T1 image was acquired on the first experimental day (TR = 817 ms, TE = 3.73 ms, 240x240x223 matrix, 1-mm isotropic resolution). All four experimental learning phases were performed in different sessions in the scanner with a BOLD echo-planar imaging sequence (TR = 2.53 s, TE = 30 ms, 96x96x46 matrix, 2.5-mm isotropic resolution). Further, phase-opposite scans (with otherwise identical acquisition parameters) were acquired for each task session to correct for distortion artifacts.

MRI preprocessing

MRI data were preprocessed with fMRI prep (Esteban et al., 2017; https://fmriprep.org/) as described below.

Anatomical data preprocessing

The T1-weighted (T1w) image was corrected for intensity non-uniformity with N4BiasFieldCorrection (Tustison et al. 2010), distributed with ANTs 2.2.0 (Avants et al. 2008), and used as T1w reference throughout the workflow. The T1w-reference was then skull-stripped with a *Nipype* implementation of the antsBrainExtraction.sh workflow (from ANTs), using OASIS30ANTs as the target template. Brain tissue segmentation of cerebrospinal fluid, white-matter, and gray-matter was performed on the brain-extracted T1w using fast (FSL 5.0.9, Zhang, Brady, and Smith 2001). Brain surfaces were reconstructed using *recon-all* (FreeSurfer 6.0.1, Dale, Fischl, and Sereno 1999), and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle (Klein et al. 2017). Volume-based spatial normalization to standard space (MNI152NLin2009cAsym) was performed through nonlinear registration with *antsRegistration* (ANTs 2.2.0), using brain-extracted versions of both the T1w reference and T1w template. The following template was selected for spatial normalization: *ICBM 152 Nonlinear Asymmetrical template version 2009c* (Fonov et al. (2009).

Functional data preprocessing

For each of the four BOLD sessions per subject, the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) were estimated before any spatiotemporal filtering using mcflirt (FSL 5.0.9, Jenkinson et al. 2002). BOLD sessions were slice-time corrected using 3dTshift from AFNI 20160207 (Cox and Hyde 1997, RRID:SCR_005927). A deformation field to correct for susceptibility distortions was estimated based on *fMRIPrep*'s *fieldmap-less* approach. The deformation field results from co-registering the BOLD reference to the same-subject's T1w-reference with its intensity inverted (Wang et al. 2017; Huntenburg 2014). Registration was performed with antsRegistration (ANTs 2.2.0), and the process regularized by constraining deformation to be nonzero only along the phase-encoding direction, and modulated with an average fieldmap template (Treiber et al. 2016). Based on the estimated susceptibility distortion, a corrected echo-planar imaging reference was calculated for a more accurate co-registration with the anatomical reference. The BOLD reference was then co-registered to the T1w reference using bbregister (FreeSurfer), which implements boundary-based registration (Greve and Fischl 2009). Co-registration was configured with six degrees of freedom. The BOLD time-series were resampled onto the following surfaces (FreeSurfer reconstruction nomenclature): *fsnative*. The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying a single, composite transform to correct for head-motion and susceptibility distortions. These resampled BOLD time-series will be referred to as *preprocessed BOLD in original space*, or just *preprocessed BOLD*. The BOLD time-series were resampled into standard space, generating a *preprocessed BOLD session in MNI152NLin2009cAsym space*. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. Several confounding time-series were calculated based on the *preprocessed BOLD*: framewise displacement (FD), DVARS and three region-wise global signals. FD was computed using two formulations following Power et al. (2014) (absolute sum of relative motions) and Jenkinson et al. (2002) (relative root mean square displacement between affines). FD and DVARS were calculated for each functional session, both using their implementations in *Nipype* (following the definitions by Power et al. 2014). The three global signals were extracted within the cerebrospinal fluid, white matter, and whole-brain masks.

Univariate analyses

We estimated activation differences between the different CS types for each experimental phase by computing 1st level contrasts of interest (e.g.: CS+>CS-), followed by a 2nd level group analysis with a FWE correction at the cluster level. Significant cluster size was estimated in a non-parametric manner using nilearn's *non_parametric_inference* function, using 10,000 permutations. Analyses were constrained within a grey matter mask.

RSA

We applied RSA to investigate the representational geometries of cues and contexts throughout the experiment (Figure 1C). Since this involved studying the neural representations during a relatively rapid event-related design, we chose to base the estimation of neural pattern similarity not on the raw BOLD data, but rather on General Linear Model (GLM) estimates of the trial-specific BOLD response, an approach known as beta-series modeling (Rissman et al., 2004; Turner et al., 2012). Specifically, we used a Least Square Separate (LSS) approach (Abdulrahman and Henson, 2012), which consists of fitting one GLM for each trial, in which the tested trial is the condition of interest while controlling for all other trials. We chose this approach over the Least Square All method, which consists of fitting a single GLM including all trials (and using each trial as a condition of interest), as LSS has been shown to be superior for dealing with collinearity, especially with fast event-related designs (Abdulrahman and Henson, 2012; Mumford et al., 2012), which applies to our paradigm.

We estimated two separate sets of LSS models per experimental phase: one set of GLMs for CS, and a distinct set of GLMs for contexts. The onset of CS was not included in the context models, and vice-versa. We reasoned that including both CS and contexts in the same LSS models was not necessary as CS type and video type were orthogonal to each other, i.e., the paradigm intrinsically controls for the potential influence of CS on context and vice-versa. Furthermore, including all CS and context events in one model may risk overfitting the data (since all events but one serve as variables of non-interest in each LSS model). However, to distinguish these two events temporally, the onset of the context was set as the beginning of each video and its offset as the presentation of the CS. In all models, to avoid the confounding effect of the US (electric shock) on the CS pattern, only unreinforced trials (i.e., not followed by a US) were used to conduct statistical analyses of neural pattern similarity at the group level. The onset of each US was also used as a regressor of no interest in the LSS models of the phases where US were presented (i.e., for fear acquisition and reversal). The six motion parameters (three translation and three rotation) and the average signal in the white matter and corticospinal fluid compartments were also added as regressors of non-interest. This LSS beta-series approach was implemented using nilearn (https://nilearn.github.io/).

Trial-wise pattern similarity of each cue or context was obtained on the beta-series as the temporal correlation between all cue trials or context trials with a searchlight approach and region of interest (ROI) approach. Raw correlation values were, in both cases, Fisher r-to-z transformed before any further analyses.

Item stability and generalization of cues

For both cues and context, we analyzed two different types of pattern similarity, i.e., item stability (within-stimulus similarity) and cue generalization (between-stimulus similarity). Item stability was defined as the average within-cue or within-context neural pattern similarity, i.e., the average neural pattern similarity between the different presentations of a given cue (out of eight possible cues per phase and in the whole experiment) or a given context (out of four possible contexts per phase, for a total of 16 contexts in the whole experiment). Thus, item stability represents the similarity of the neural representation of an item to other representations of this same item (Xue, 2018), or the consistency of neural activity across repetitions (Sommer et al., 2022).

On the other hand, cue generalization was defined as the average neural pattern similarity between different exemplars of a same CS type (e.g., the similarity between the two different CS-+ items during reversal learning). Thus, while item stability provides information about the stability of the neural representation of one particular item, cue generalization expresses the formation of a higher-order association between different exemplars of the same valence category (see Visser et al., 2013 for a similar approach). In the case of contexts, cue generalization was defined as the average neural pattern similarity between the different videos presented in a given experimental phase.

RSA of context specificity

Specifically for contexts, we computed the representational specificity of contexts in each experimental phase, which was defined as the difference of the average similarity of the different presentations of the same contexts (i.e., within-context similarity, or context stability) with the average similarity of the different presentations of different contexts (between-context similarity, or context generalization), separately for each experimental phase. In other words, context specificity controls for the stability of a particular stimulus (i.e., context video) with the generalization between distinct stimuli (all other contexts shown in an experimental phase), with higher context specificity entailing more distinct representations of contexts in each experimental phase. We then compared the context specificity maps between phases, in order to assess the effect of experimental manipulation (acquisition, reversal, and test phases) on context specificity.

Searchlight approach

The different types of similarity analyses described above were implemented in a searchlight approach. Pattern similarity was estimated at the voxel level using the searchlight algorithm as implemented in brainIAK (https://brainiak.org/). A square with a radius of twice the voxel size (i.e., 5-mm radius) was used with each brain voxel as the center to estimate the average pattern similarity within the searchlight. Voxels were included only if at least 50% of their surrounding voxels were included in the brain mask. A pattern similarity formula specific to each type of pattern similarity (e.g., item stability and cue generalization for CS or contexts) was used. Hypothesis testing was performed by comparing the obtained Fisher r-to-z-transformed correlation maps of the conditions of interest (e.g., CS+ cue generalization vs. CS- cue generalization). Significance corresponding to the contrast between conditions of the maps of interest was estimated using non-parametric permutation tests at the cluster level, with 10,000 permutations used to estimate significant cluster size. Analyses were restricted within a gray matter mask.

ROI-based RSA

Additionally, pattern similarity analyses were performed at the ROI level. We thresholded the statistical maps of the searchlight analyses to only retain (corrected) significant clusters. We used nilearn's *connected_region* function to define individual ROIs statistical maps, using a minimum ROI size of 1500mm^3. These ROI masks in MNI space were used to estimate the average neural pattern similarity within each ROI, defined as the correlation between the BOLD response of all trials. Raw correlation values were then Fisher r-to-z transformed. The effects of experimental phase and type of neural pattern similarity (within/between) were assessed with LME models using the *lme4* package in R. Significance was assessed with the Satterthwaite method for estimating degrees of freedom using maximum likelihood. All statistical analyses were corrected for multiple comparisons using False Discovery Rate (FDR) unless indicated otherwise.

3. Results

Behavioral results

We first examined the trial-wise US expectancy ratings across experimental phases. A linear mixed effects (LME) model with "CS type" and "experimental phase" as fixed effects and "participant" as a random effect revealed significant effects of CS type (F(1816.6, 605.54)= 479.35, p<0.0001) and experimental phase (F(476.3, 158.78)=125.6, p <0.001) as well as a significant interaction (F(334.8, 37.2)=29.45, p<0.001), showing that both CS type and experimental phases affected US expectancy (Figure 2A). Post-hoc paired Wilcoxon tests (Bonferroni-corrected) showed that ratings to all CS types were significantly different from each other across all the experimental phases (CS++ > CS+- > CS-+ > CS--; all p<0.01) except during fear acquisition in which CS++ and CS+- cues on the one hand, and CS-+ and CS-- cues on the other hand were equivalent, as expected at this stage. Post-hoc pairwise comparisons between experimental phases across all CS types were significant as well (reversal > acquisition > test$_{new}$ > test$_{old}$; all p<0.0001), and fear reversal and acquisition were the experimental phases with the highest US expectancy. The CS type with the highest US expectancy was, as expected, CS++.

To examine the interaction between US expectancy and CS type, we conducted post-hoc Wilcoxon tests (Bonferroni-corrected), which revealed different patterns of CS differences between experimental phases (Supplementary Table 1).

Activation of fear network by cues signaling current and prior threats

Next, we assessed activity differences between CS types during each experimental phase (Figure 2B). During acquisition, the CS+ > CS- contrast showed significantly increased BOLD activity in several clusters across the fear network, such as the dACC, superior frontal gyrus, caudate nucleus, and middle temporal gyrus (Figure 2Bi, in line with previous work (e.g., Fullana et al., 2016). The opposite contrast CS- > CS+ showed no significant clusters.

During reversal, a contrast of current valence, i.e., (CS++ and CS-+) > (CS+- and CS--) showed activation patterns similar to those during acquisition, spanning across the fear network (Figure 2Bii). We then contrasted currently threatening and safe cues depending on their previous valence, i.e., (CS++ > CS+-) > (CS-+ > CS--), which also revealed activation in the fear learning network, although to a lesser extent (Figure 2Biii). This result may reflect the impact of the lingering Pavlovian trace (remaining from acquisition) and/or the time required to learn contingency changes during reversal.

During the two test phases, none of the contrasts between CS types (CS++ > CS--, CS-+ > CS--, CS+- > CS--) revealed any significant activity differences. Thus, BOLD responses were similar for all CS types in the absence of a US, even though differences in US expectancy ratings were observed at the behavioral level. This further underlines the necessity for an analysis of representational patterns rather than mere univariate activity differences.

Generalized representations of threat cues during acquisition and reversal

We thus focused our analyses on the representational geometry of cues across experimental phases. We examined the effect of cue type on two distinct representational properties, cue generalization (between-cue similarity) and item stability (within-cue similarity), using a whole-brain searchlight approach (Figure 3A).

During fear acquisition, we again combined CS++ and CS+- cues (both followed by a US in 50% of trials) into a common CS+ category, and CS-- and CS-+ cues (never followed by a US) into a common CS- category. We found that item stability did not differ between CS+ and CS- cues. Importantly, however, cue generalization was significantly higher for CS+ compared to CS- cues in several clusters across the fear network, with a pattern reminiscent of the results during the corresponding univariate analyses. Thus, the dACC, superior frontal gyrus, caudate nucleus, and insula were among the regions showing higher cue generalization of CS+ compared to CS- cues (Figure 3C; Supplementary Table 2). This suggests the formation of a higher-order association (i.e., a category-level stability) between threatening cues and less so between safe cues during fear acquisition. The opposite contrast (CS- > CS+) did not reveal any significant effects.

Distinct functional roles, spatial distributions, and subsequent persistence of item stability and cue generalization during reversal

Next, we compared item stability and cue generalization between the four CS types during fear reversal (Figure 3B-D). We first compared CS++ and CS-- cues, corresponding to the CS+ vs. CS- contrast during fear acquisition (Figure 3B). Again, we observed higher cue generalization for CS++ compared to CS-- cues in the dACC, but not in any of the other regions where cue generalization effects were observed during acquisition. Comparing the cue generalization of all CS cues that are threatening in reversal (CS++ & CS-+) to the ones that are not (CS+- & CS--) revealed similar results to fear acquisition, with increased cue generalization across fear learning network regions including the dACC, superior frontal gyrus (SFG), medial temporal gyrus, and inferior frontal gyrus (IFG) (Figure 3C). Item stability, in line with results from fear acquisition, did not differ between currently threatening and non-threatening cues. Moreover, we found no significant clusters when comparing item stability between CS++ vs. CS-- cues.

We then investigated representational effects of contingency changes and compared cues that changed their contingency between acquisition and reversal (CS+- and CS-+) to cues with consistent contingencies (CS-- and CS++), resulting in the ‘change vs. no-change’ contrast, i.e., (CS+- & CS-+) > (CS++ & CS--). Interestingly, item stability was significantly higher for changing than consistent cues in the precuneus and IFG, i.e., in regions that overlapped with those showing higher cue generalization for threatening cues (Supplementary Table 3; Figure 3Biii). Conversely, this contrast did not reveal any differences in cue generalization.

We next investigated item stability and cue generalization during test$_{new}$ and test$_{old}$, i.e., when USs are absent for all CS types. In order to understand the impact of prior contingencies – i.e., of lingering Pavlovian traces – on test$_{new}$ and test$_{old}$, we examined the contrast between CS++, CS-+, and CS+- with CS-- (safe baseline), as well as the contrast between CS-+ and CS+- with CS++ (unsafe baseline) in these two phases.

Cue generalization did not differ between CS types during either test$_{new}$ or test$_{old}$. However, item stability was higher for CS+- vs. CS++ cues in a middle temporal cluster during test$_{new}$ (Figure 3E), and for CS++ compared to CS-- cues in an inferior temporal cluster during test$_{old}$ (Figure 3F; Supplementary Table 4).

To summarize, cue generalization and item stability showed an interesting dissociation during reversal, with higher cue generalization for threatening vs. non-threatening cues, and higher item stability for changing vs. consistent cues. This suggests that these two representational properties could capture distinct aspects of contingency learning; the threatening (vs. safe) nature of cues increased cue generalization, while a changing (vs. consistent) nature of contingencies enhanced item stability. We also found that item stability but not cue generalization effects persisted in the absence of a US.

Memory traces from previous learning phases compete for reinstatement during test

Fear extinction is commonly described as an inhibitory process, in which a new safety memory trace is created and competes with the previous threat memory trace that is concurrently inhibited (Lebois et al., 2019; Santini, 2008; Szeska et al., 2020). In our paradigm, the memory traces formed during acquisition and reversal might compete for reinstatement in particular during the test$_{old}$ phase, because both fear and reversal memories may reoccur during this phase due to context overlap. We compared the magnitude of reinstatement effects between acquisition and reversal during test$_{old}$, by using an LME with experimental phase, CS type, and their interaction as predictors, and reinstatement as the predicted variable. Reinstatement was estimated by comparing either the similarities of identical items across phases (item reinstatement) or the similarities of different items from one cue type across phases (generalized reinstatement). We extracted these reinstatement values from significant clusters observed in our previous searchlight analyses (Figure 4A), and correlated them across participants (see Graner et al., 2020 for a similar approach).

We observed a significant effect of experimental phases on item reinstatement in IFG (F(2,253)=5.50, p<0.01) (Figure 4Bi). Post-hoc Wilcoxon t-tests showed that reversal-test$_{old}$ item reinstatement was significantly higher than acquisition-test$_{old}$ item reinstatement (t(253)=-3.01, p<0.01) and test$_{new}$-test$_{old}$ item reinstatement (t(253)=-2.7, p<0.05). In addition, we observed a significant effect of experimental phases on generalized reinstatement in dmPFC (F(2,259)=4.01, p<0.05) (Figure 4Bii), where acquisition-test$_{old}$ generalized reinstatement was higher than test$_{new}$-test$_{old}$ generalized reinstatement (t(259)=2.96, p<0.05). In summary, during test$_{old}$, we observed prominent reinstatement of item-specific reversal memory traces in IFG and of generalized acquisition memory traces in dmPFC, suggesting that memories from these two phases tend to come back in different representational formats and in dissociable brain regions.

Context specificity increases during reversal in prefrontal cortex and predicts the reoccurrence of fear memory traces

We followed our analyses of cue generalization, item stability, and reinstatement with an analysis of the neural representation of contexts across experimental phases. Given that memories built during extinction learning tend to be more context-specific than those acquired during initial fear learning (Maren et al., 2013), we established a measure of context specificity, namely the difference between neural representations of same vs. different contexts in each phase. We also compared this measure across experimental phases and related it to the representational geometries of cues (Figure 5A).

As hypothesized, we found that context specificity during reversal was significantly higher than it was during acquisition, an effect that occurred in a cluster including both dorsomedial PFC and lateral PFC (i.e., superior frontal gyrus), areas known to be involved in contextual processing (Maren et al., 2013; Figure 5B, Supplementary Table 5). Context specificity did not differ between the other experimental phases.

Previous research has suggested that higher context specificity during extinction learning predicts a more pronounced reoccurrence of fear memories (LaBar and Phelps, 2005; Milad et al., 2005; Vansteenwegen et al., 2005; Neumann, 2006; Navarro-Sanchez et al., 2024). To investigate this hypothesis, we compared the neural representations of cues between experimental phases and correlated the magnitude of reinstatement effects with the increase in context specificity from acquisition to reversal (across participants). We extracted subject-wise measures of context specificity from the PFC cluster in Figure 4B and compared them to both item reinstatement and generalized reinstatement. We extracted these reinstatement values from significant clusters resulting from our previous searchlight analyses (Figure 4A, Figure 5C).

We focused our analyses on cues that changed their contingencies (i.e., CS-+ and CS+-), which were expected to reveal differential reinstatement of acquisition vs. reversal memory traces. We used LME models with “context specificity” and “cue type” as predictors and “item reinstatement” or “generalized reinstatement” as the dependent variable (Figure 5C). Correction for multiple comparisons was done (FDR) within each item reinstatement/generalized reinstatement pair of each ROI.

We tested whether increased context specificity during reversal predicted the reinstatement of acquisition memory traces during test$_{old}$ (the test phase with the acquisition/reversal contexts). We found that an interaction between context specificity and cue type predicted generalized reinstatement of acquisition memory traces in both ACC/SFG ($t(22)=6.25$, $p<0.05$) and precuneus ($t(22)=4.89$, $p<0.01$) (Figure 5Di). In the precuneus, higher context specificity during reversal predicted more generalized reinstatement of the initially non-threatening cues (CS-+), i.e., a cue type that is safe during both acquisition and test, as compared to initially threatening cues (CS-+). Reversely, in the ACC/SFG, higher context specificity during reversal predicted more generalized reinstatement of the initially threatening cues (CS+-) than of the initially safe cues (CS-+).

In addition, higher context specificity during reversal predicted higher item reinstatement of reversal memory traces for CS-+ than CS+- cues in the dmPFC (t(22)=5.56, p<0.05). Thus, similar to reinstatement in ACC/SFG, higher context specificity again favored reinstatement of memory traces of threatening over safe cues, even though these memory traces were now from the reversal rather than the acquisition phase (Figure 5Dii).

As a control, we also analyzed reinstatement during the test$_{new}$ phase. We tested whether increased context specificity during reversal predicted the reinstatement of acquisition memory traces during this phase with new contexts. We found that an interaction between context specificity and CS type predicted the reinstatement of acquisition memory traces in middle temporal gyrus (t(22)=2.51, p<0.05) (Figure 5Diii). Specifically, higher reversal context specificity predicted more item reinstatement of CS-+ cues, i.e., cues that were safe during acquisition, than of the initially threatening CS+- cues.

Together, these results indicate more specific context representations during reversal than acquisition (Figure 5B) and show that acquisition memory traces are predominantly reinstatement in a generalized format, while reversal memory traces are reinstated at the level of individual items (Figure 4B). Perhaps most interestingly, they suggest a possible mechanism for the previously observed impact of extinction contexts on fear renewal, because higher levels of context specificity during reversal favored the reinstatement of threat memory traces in areas of the fear network (ACC and dmPFC; Figure 5Di left and 5Dii). These effects were not observed in the precuneus (Figure 5Di right) and for reinstatement during new contexts (Figure 5Diii).

4. Discussion

The present study investigated the dynamic changes in neural representations of CS cues and contexts during acquisition, reversal, test in new contexts (test$_{new}$), and test in previous acquisition/reversal contexts (test$_{old}$). Our main findings demonstrate distinct representational properties of CS cues and contexts during these different phases, suggesting that representational geometries reflect the fate of memory traces. We found that (1) cue generalization and item stability play complementary roles during initial fear learning and reversal, by being associated with threatening-vs-safe cues and changing-vs-consistent cues, respectively; (2) during test$_{new}$ and test$_{old}$, differences of cue generalization between CS types disappear, while some differences of item stability remain; (3) context representations become more specific following contingency changes during reversal learning; and (4) the context specificity during reversal predicts the reinstatement of fear memories during subsequent test, providing a mechanistic basis for clinically relevant phenomena such as renewal. These results offer new insights into the regional distributions, representational geometries, and functional relevance of cues and contexts across distinct stages of fear learning, opening new avenues of understanding fear-guided behavior.

Complementary representational properties during initial fear learning

Our results demonstrate how different representational properties of CS cues are associated with distinct aspects of fear learning, i.e., cue generalization with a threatening vs. safe nature of the CS, and item stability with a changing vs. consistent nature of the CS. Consistent with previous studies (Visser et al., 2011; 2013), we found that cue generalization was greater for CS+ than for CS- cues during fear acquisition in regions of the fear network (e.g., ACC) and the salience network. This suggests that fear acquisition leads to the formation of a higher-order association between different reinforced cues, but less so between unreinforced ones. This category-level learning could allow for efficient threat detection and generalization, an adaptive behavior in potentially dangerous environments. Moreover, previous studies showed that the role of cue generalization in the coding of threat extends beyond fear conditioning, as shown by Dunsmoor et al. (2014) who found enhanced memory consolidation of items sharing conceptual similarity with threat-associated stimuli.

In stark contrast, item stability of CS cues, i.e., their within-stimulus similarity across repetitions, did not differ between CS+ and CS- cues during fear acquisition. This indicates that while acquisition induces a category-level representation for reinforced cues, it does not differentially modify the item-level representations of CS+ as compared to CS- cues. By contrast, item stability was particularly sensitive to changes in CS valence between experimental phases, suggesting that it plays a crucial role in tracking and updating the specific threat associations of individual stimuli. Indeed, in fear reversal, while cue generalization remained greater for reinforced compared to unreinforced cues (CS--), mirroring the pattern observed during acquisition, item stability specifically increased for cues that changed valence between acquisition and reversal (CS-+ and CS+-). This finding suggests that when the contingencies change, the participants might focus more on the individual properties of the cues to interpret the new contingencies, leading them to fine-tune their representations. Indeed, item stability has been linked to successful memory encoding and retrieval by several studies (Xue et al., 2010; LaRocque et al., 2013; Zheng et al., 2018). Moreover, neural correlates of item stability have been reported in regions of the episodic memory network, such as the IFG and the precuneus (Xue et al., 2010), where we found a significant effect of item stability during reversal. Therefore, item stability might hence be more akin to an episodic-like type of learning, while cue generalization might be more reflective of category-level learning (Visser et al., 2013).

Previous findings by Visser et al. (2011, 2013) demonstrate distinct learning curves for item stability between CS+ and CS- cues from trial to trial. This discrepancy could be caused by methodological differences, as our study focused on session-wise differences of item stability for each cue type and not on trial-by-trial differences. In line with our conclusions, however, Visser et al. (2013) found that item stability was increased for subsequently remembered cues, while cue generalization was associated with the later behavioral expression of fear memory. Overall, the increased item stability during reversal of the items that change contingency could reflect a process of stabilizing the new valence at the item level, as the change of contingency may lead to the temporary representation of individual items as ‘categories’ themselves, without being subsumed yet into the generalized representations encompassing multiple different items sharing the same valence. This dual representational signature may allow for both efficient threat detection (via category representations) and flexible updating of individual stimulus associations (via item-specific representations).

Dissolution of cue generalization and item stability in the absence of US

During the test phases, we did not observe any differences in cue generalization between cue types. Further analyses comparing cue generalization values between learning phases revealed a decrease in cue generalization during the test phases compared to acquisition and reversal, mostly affecting previously threatening cues (Supplementary Figure S1). We also observed a decrease in item stability from acquisition/reversal to the test phases, which affected cues that changed valence (CS-+, CS+-) more than cues with consistent valence (CS++, CS--) (Supplementary Figure S1). However, some differences in item stability remained during test$_{new}$ (higher for CS+- vs CS++ in the middle temporal gyrus) and test$_{old}$ (higher for CS++ vs CS-- in the inferior temporal cortex) (Figure 3C-D).

These findings suggest that the disappearance of threat during the test phases may involve two concurrent processes: (1) An unlearning of generalized threat representations, evidenced by the global decrease in cue generalization; and (2) a partial unlearning of item-level representations, particularly for cues with changing contingencies, reflected in diminished item stability. Interestingly, these effects occurred during the test phases rather than during reversal, suggesting that they are driven by the absence of the US rather than by the contingency change. During reversal, the continued presence of the US, albeit with a different contingency, may still benefit from generalized representations at the item and category levels. Contrastingly, the complete absence of the US during the test phases may promote a differentiation of CS representations, as the need for generalization diminishes. This finding highlights the importance of considering the specific reinforcement history of cues for understanding the dynamics of fear representations.

Reinstatement of item representations during test is weaker for fear extinction

Our results showed a differentiation of fear memories during the test phases, both at the item and category level. Interestingly, we found that the reinstatement of representations from the first test phase during the immediately following second test phase was weaker compared to the reinstatement of acquisition and reversal memory traces that had been formed on a different day (Figure 4E). This may be explained by the greater differentiation of cue representations during test$_{new}$; the more differentiated the representations, the less likely they are to be reinstated subsequently. The weaker reinstatement of memories from a phase without any US, compared to memories from acquisition and reversal phases with US, may contribute to the challenges of preventing relapse in anxiety disorders (Vervliet et al., 2013). If extinction learning results in less stable and less generalizable safety representations, individuals may remain vulnerable to the return of fear once they return to previous contexts (Boschen et al., 2009).

Increased specificity of context representations following contingency changes

Our analysis of context representations revealed an increased specificity of context coding during reversal compared to initial acquisition. This suggests that the brain may allocate more resources to representing contextual details when contingencies are changing, by potentially facilitating the adaptive updating of contingencies against a more stable contextual backdrop.

The dorsomedial PFC, including the superior frontal gyrus and ACC, have emerged from our analyses as key regions exhibiting higher context specificity in reversal learning. Given their roles in attentional control (Dosenbach et al., 2007) and conflict monitoring (Stevens et al., 2011), the dmPFC's involvement may reflect the increased attentional and control demands induced by changing contingencies. Computationally, the more precise contextual coding in these regions during reversal could serve to disambiguate cues of changing contingencies, supporting the formation of new context-dependent associations (Xu & Südhof, 2013). Our findings extend prior work on the importance of hippocampus and mPFC in representing context during fear learning and extinction (Maren et al., 2013), as these regions could dynamically adjust their representational specificity in response to a change in environmental demands.

Context specificity is associated with reinstatement of fear memory traces

The amount of reinstatement during test$_{old}$ was related to the increase in context specificity from acquisition to reversal. We quantified this increase in specificity in the dmPFC cluster identified in the previous analysis and correlated it with two measures of reinstatement: (1) item reinstatement, reflecting the similarity of individual cue representations between phases; and (2) generalized reinstatement, capturing the similarity of cue representations among their CS category.

For regions involved in threat processing, such as the ACC/SFG, higher context specificity predicted stronger generalized reinstatement of representations of previously threatening cues (CS+-) from acquisition to the test phase. This suggests that for these cues, the more distinct the contextual coding during reversal, the more strongly the original fear memory trace resurfaced, likely reflecting a return of fear (Figure 5D). Contrastingly, for areas implicated in cue-specific processing that could reflect more episodic-like learning, such as the precuneus (Cavanna & Trimble, 2006), context specificity was associated with enhanced generalized reinstatement for cues with consistent meanings across phases (e.g., CS+- cues from reversal to test). Regarding item reinstatement, the dmPFC behaved similarly to the ACC/SFG, with stronger item reinstatement of previously threatening cues (CS-+ from reversal to test), while the MTG showed a pattern similar to the precuneus, with stronger item reinstatement for cues of consistent meanings across phases.

These findings highlight the region-, phase- and cue-specific effects of contexts on the reinstatement of cue representations. In threat-responsive regions, context specificity may promote the resurgence of generalized threat representations, in line with notions of renewal and spontaneous recovery of fear (Maren et al., 2013). Conversely, in episodic learning regions, contextual coding may support the reactivation of representations when meanings are maintained, reflecting memory stability. Together, these results suggest a critical role of context representations in modulating the balance between generalization and specificity of fear memories over time.

Limitations and future directions

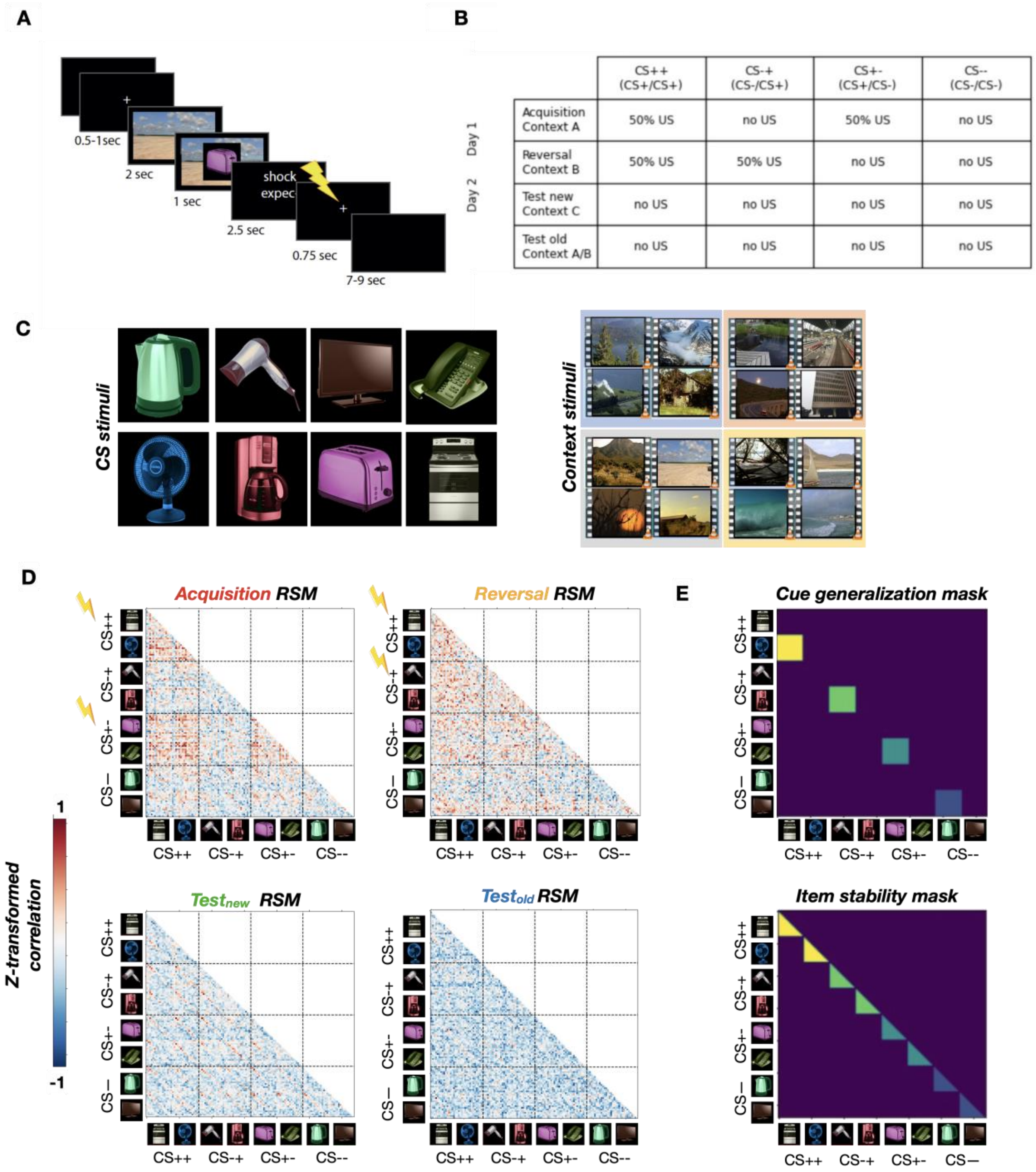
While our study provides novel insights into the changes of neural representations across the different stages of fear learning, reversal, and test, several limitations should be noted. First, our sample size was relatively small, and future studies with larger samples will be needed to replicate and extend our findings. Second, while we examined the spatial patterns of neural activity using RSA, we did not assess potential changes in the temporal dynamics of these patterns. Several studies have highlighted the importance of considering temporal information in understanding the neural mechanisms of fear learning (Bach et al., 2011; Visser et al., 2013; Sperl et al., 2021). Integrating spatial and temporal pattern analysis in future studies could provide a more comprehensive overview of how fear representations evolve over time. Moreover, further examining the role of context manipulation, by using more classical approaches where only one context is presented per phase, could extend and generalize our current findings. Finally, applying our approach to clinical populations could yield important insights into the neural mechanisms underlying the overgeneralization of fear, and the impaired contextual regulation of fear responses in psychiatric disorders.

Conclusion

Our study reveals the changes in neural representations of conditioned stimuli and contexts across fear learning phases. Cue generalization and item stability play complementary roles in fear acquisition, reversal, and test, by capturing the formation of threat-related categories and the updating of the contingency of individual stimulus representations, respectively. Phases devoid of US cues lead to a differentiation (or dissolution) of both category- and item-level representations. Context specificity in the prefrontal cortex modulates the persistence of fear memories, with region-specific reinstatement effects. These findings provide insights into the representational dynamics underlying fear learning and extinction, demonstrating the interplay between cue- and context-based representations in shaping the formation, updating, and reinstatement of fear memories. Understanding these mechanisms might help optimize interventions targeting pathological fear in anxiety disorders. Future research should extend these findings to clinical populations and investigate the identified representational properties as biomarkers for assessing the effectiveness of extinction-based therapies.

Acknowledgments

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through grant SFB 1280 (projects A02 and A19) project number 316803389.



		CS++ (CS+/CS+)	CS-+ (CS-/CS+)	CS+- (CS+/CS-)	CS-- (CS-/CS-)
Day 1	Acquisition Context A	50% US	no US	50% US	no US
	Reversal Context B	50% US	50% US	no US	no US
Day 2	Test new Context C	no US	no US	no US	no US
	Test old Context A/B	no US	no US	no US	no US

Figure 1. Overview of the paradigm and analysis approach. A: Example structure of a trial. Each trial comprises the presentation of a context video, cue, and US expectancy rating. Electric shocks (US) are administered in reinforced trials during acquisition (following CS++ and CS+- cues) and reversal (following CS++ and CS-+ cues), with reinforcement rates of 50%. B: Paradigm structure with four different experimental phases (rows) and four different cue types (columns). Each cue type consists of two possible items. C:

CS items (left) and context videos (right). Each color indicates a set of four thematically-related context videos. Different sets are used across phases (see Table in B). D: Representational Similarity Matrices (RSMs) for each experimental phase, shown here from the dorsal ACC for illustrative purposes. Lightning images represent reinforced cue types in the different learning phases. Representations of threatening cues are more similar to each other (warmer colors), reflecting cue generalization. E: Top: Cue generalization mask for the RSA matrices estimated within each searchlight. The mask is superimposed on the RSMs (shown in C) to compute the average similarity between the different cues of each CS type (different colors). Average cue generalization values are then compared between CS types. Bottom: Item stability mask estimated within each searchlight. The mask is superimposed on the RSMs to compute the average similarity across trials of each cue, separately for each CS type (different colors). Average item stability values are then compared between CS types.

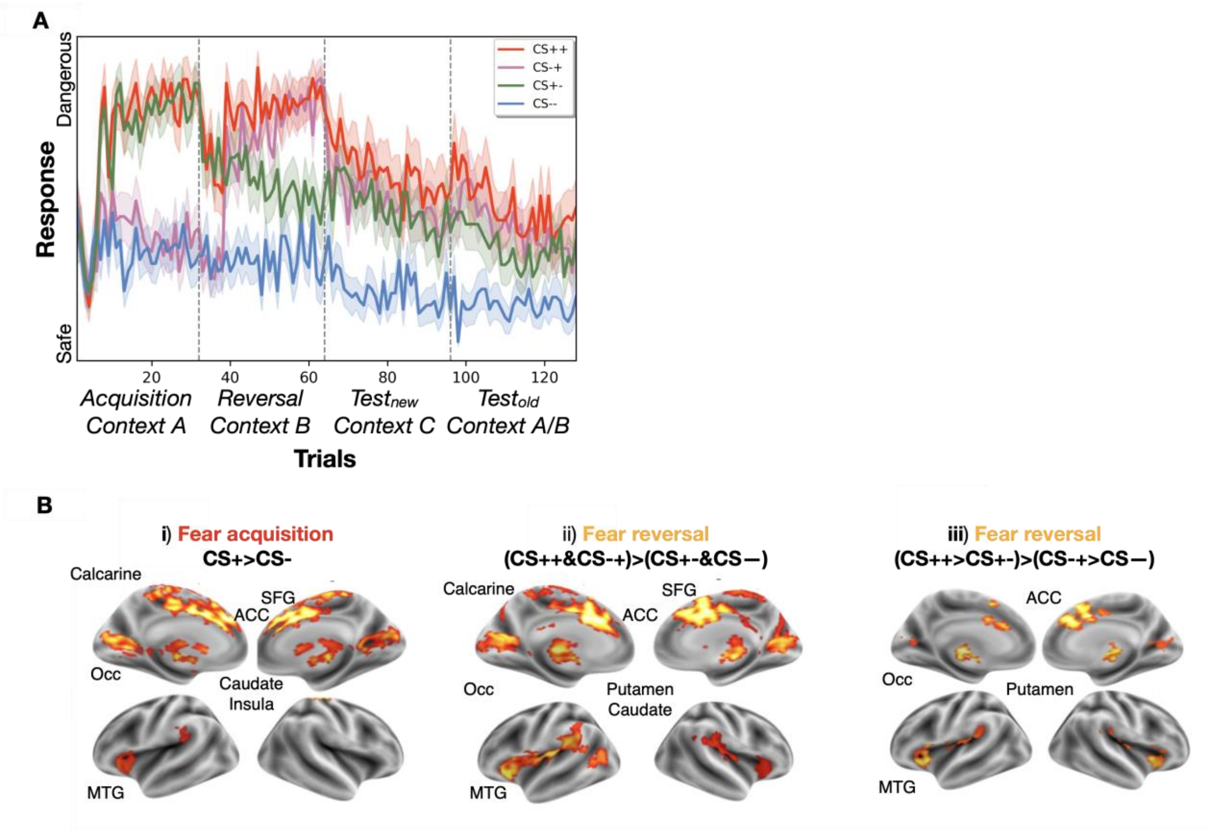


Figure 2. A: US expectancy ratings and univariate activity difference between cue types across experimental phases. Dotted lines separate the four experimental phases. Participants quickly learned the contingencies of each cue type and their changes across the experimental phases. B: Univariate activation results. Significant second-level results are shown for different contrasts in the different experimental phases. Significance was assessed at the cluster level with 10k permutations ($p_{uncorr}<0.001$).

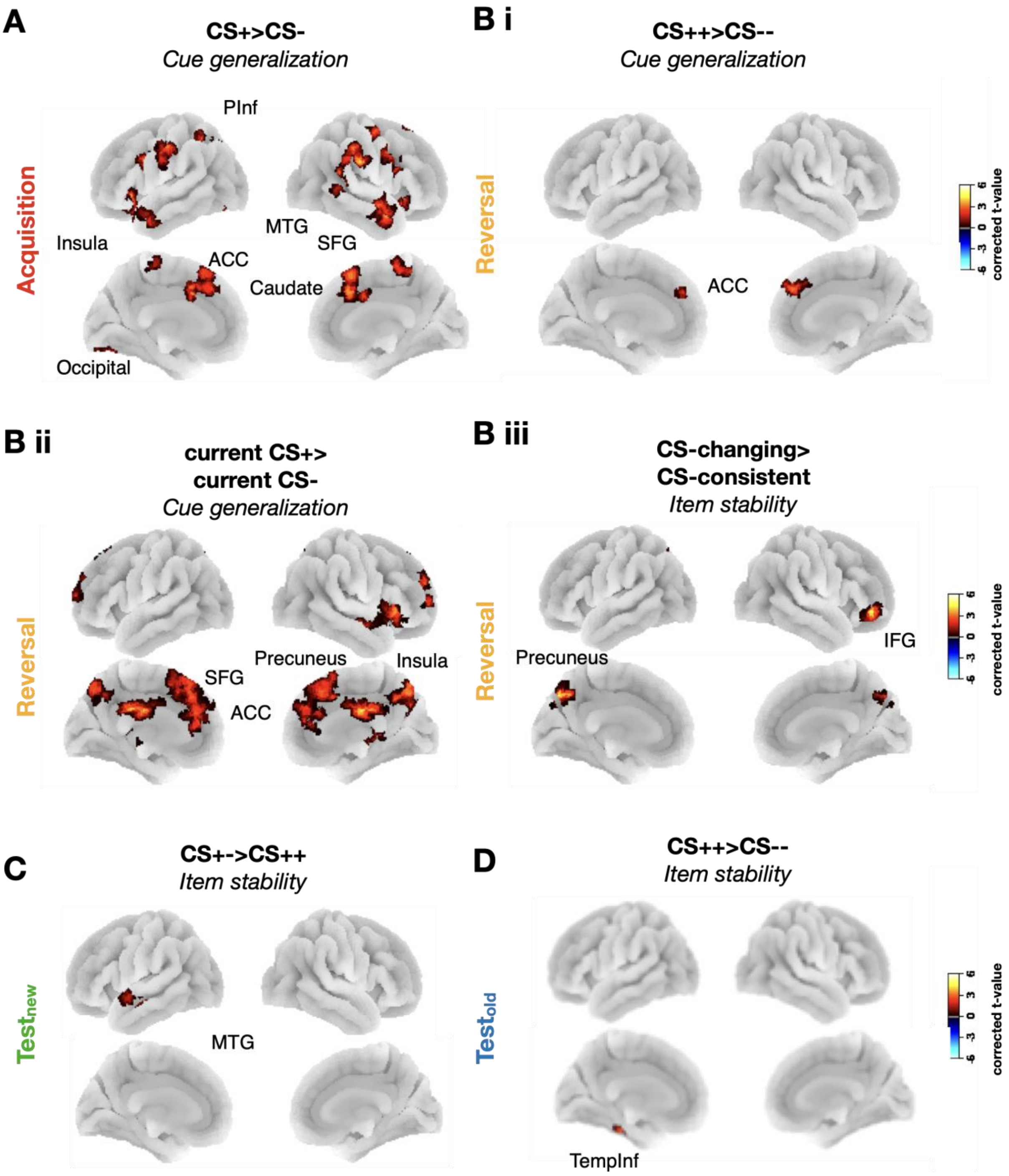


Figure 3. Enhanced cue generalization and item stability of threat cues. A: Cue representations during acquisition showing higher cue generalization of CS+ than CS- cues. No differences of item stability were found. B: Cue representations during reversal. Bi: Higher cue generalization of CS++ than CS-- cues. Bii: Higher cue generalization of currently threatening than non-threatening cues i.e., (CS-+ & CS++) > (CS+- & CS--). Biii: Higher item stability of cues with changing valence than cues with consistent valence, i.e.,

(CS-+ & CS+-) > (CS++ & CS--). C: Cue representations during test$_{new}$ showing higher item stability of previously safe cues vs. previously always threatening cues (CS+-) > (CS++). D: Cue representations during test$_{old}$ showing higher item stability of ‘previously always threatening’ vs. ‘previously never threatening’ cues (CS++) > (CS--). All plots depict t-values from searchlight analyses within family-wise error-corrected clusters (uncorrected p<0.001, corrected p<0.05) with 10k permutations.

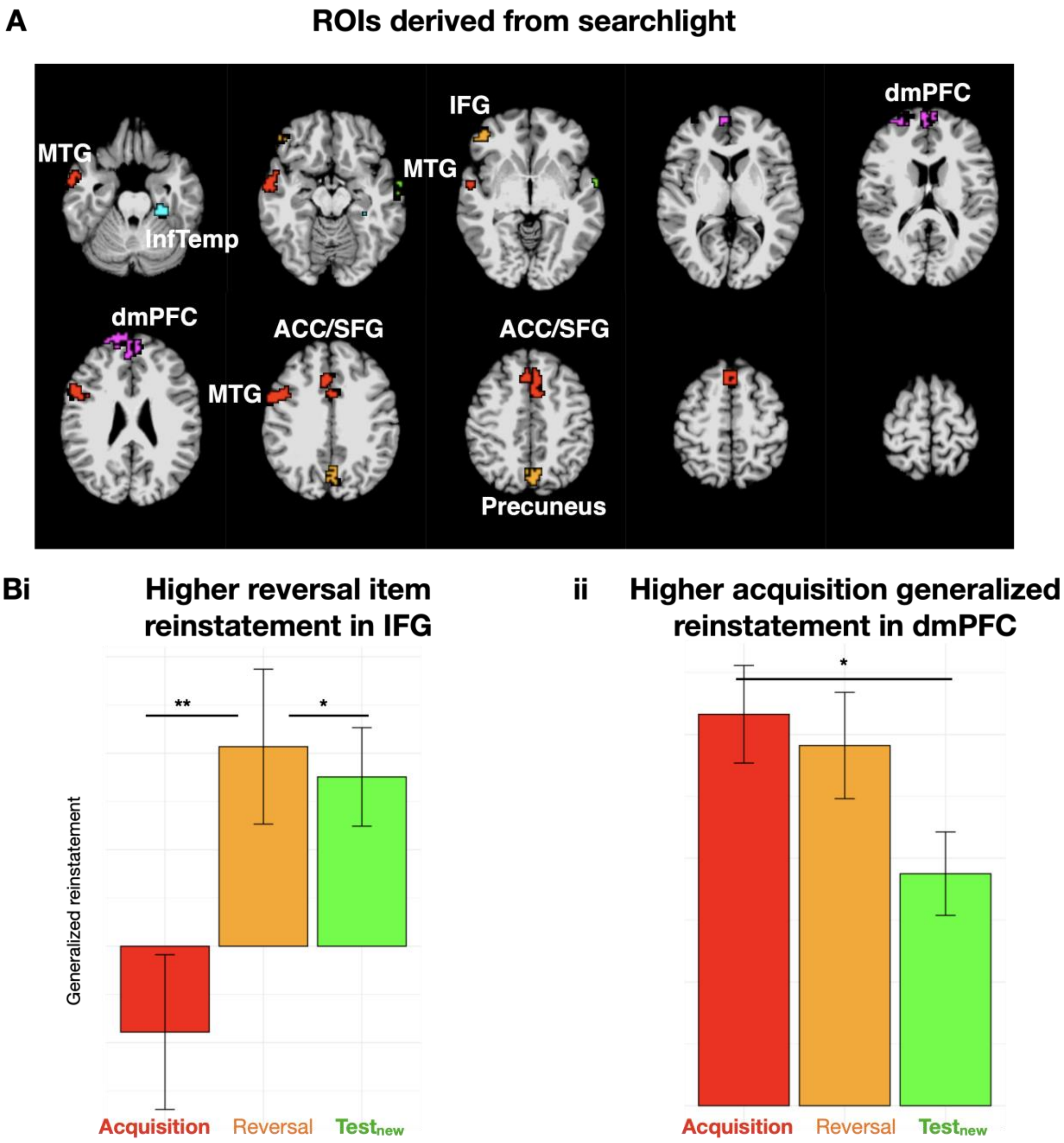


Figure 4. Different reinstatement patterns are observed for the previous experimental phases during test$_{old}$. A: ROIs derived from the previous searchlight analyses (see Figure 3), by extracting the significant clusters from the previous statistical analyses. ROIs are color-coded depending on the experimental phase they are derived from: red for acquisition, orange for reversal, green for testnew, blue for test$_{old}$. When several ROIs overlapped, only the ROI with the bigger voxel size was included in the analyses. MTG: Middle Temporal Gyrus. InfTemp: Inferior Temporal Gyrus. IFG: Inferior Frontal Gyrus. dmPFC: dorsomedial Prefrontal Cortex. ACC: Anterior cingulate

cortex. SFG: Superior Frontal Gyrus. B: Reinstatement during test$_{old}$ differed between experimental phases, such that: (Bi) in IFG, item reinstatement was higher for memory traces from reversal compared to those from acquisition and test$_{new}$; and (Bii) in dmPFC, generalized reinstatement was higher for memory traces from acquisition compared to those from test$_{new}$.

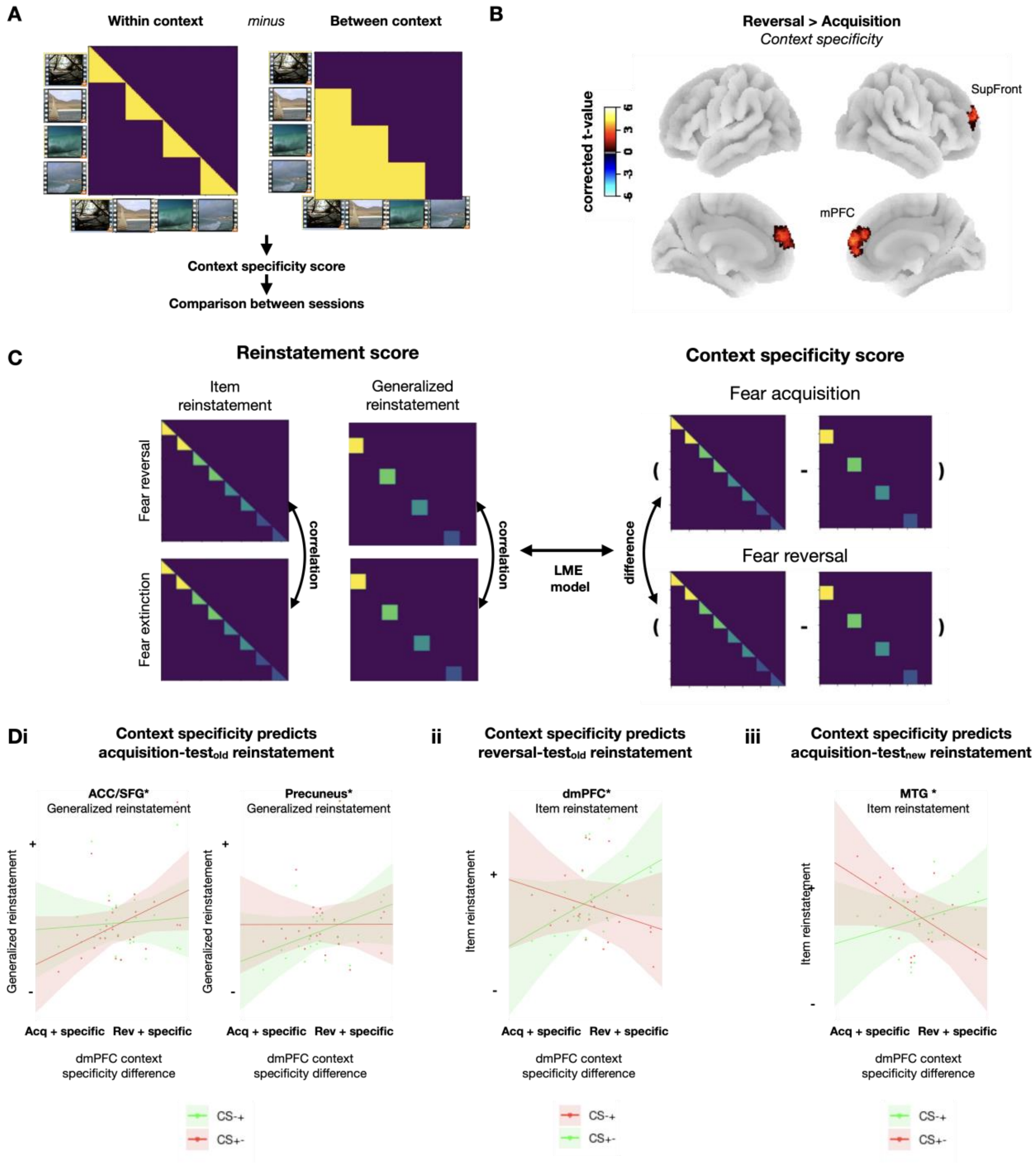


Figure 5. Context specificity during reversal and its role for reinstatement of fear memory traces. A: Calculation of context specificity as the difference of within-context similarity and between-context similarity. B: Difference in context specificity between acquisition and reversal. Positive values (in red) indicate higher context specificity in reversal. C: Calculation of item reinstatement and generalized reinstatement (similarities of item representations across different phases; left) and context specificity (difference between acquisition and reversal; right). An LME model was used to predict

these reinstatement measures by the interaction of context specificity and CS types. D: Higher context specificity during reversal predicted reinstatement during test$_{old}$, as a function of CS type: (Di) Higher reversal context specificity predicted more pronounced generalized reinstatement of CS+- vs. CS-+ acquisition memory traces in ACC/SFG (left), and reversely, more pronounced generalized reinstatement of CS-+ vs. CS+- acquisition memory traces in precuneus (right). CS+-, which is threatening in acquisition, is shown in red, and CS-+, which is not threatening in acquisition, is shown in green. (Dii) Higher reversal context specificity predicted more pronounced item reinstatement of CS-+ than CS+- reversal memory traces in dmPFC. CS-+, which is threatening in acquisition, is shown in red, and CS+-, which is not threatening in acquisition, is shown in green. (Diii) Higher reversal context specificity also predicted more pronounced item reinstatement of CS-+ (safe during acquisition; in green) than CS+- (threatening during acquisition; in red) memory traces from reversal in MTG during test$_{new}$.

References

Abdulrahman, H., & Henson, R. N. (2016). Effect of trial-to-trial variability on optimal event-related fMRI design: Implications for Beta-series correlation and multi-voxel pattern analysis. NeuroImage, 125, 756-766.

Avants, B. B., Epstein, C. L., Grossman, M., & Gee, J. C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. Medical Image Analysis, 12(1), 26–41. https://doi.org/10.1016/j.media.2007.06.004

Bach, D. R., Weiskopf, N., & Dolan, R. J. (2011). A stable sparse fear memory trace in human amygdala. The Journal of Neuroscience: The Official Journal of the Society for Neuroscience, 31(25), 9383-9389. https://doi.org/10.1523/JNEUROSCI.1524-11.2011

Beckers, T., & Kindt, M. (2017). Memory reconsolidation interference as an emerging treatment for emotional disorders: strengths, limitations, challenges, and opportunities. Annual review of clinical psychology, 13(1), 99-121.

Bierbrauer, A., Fellner, M. C., Heinen, R., Wolf, O. T., & Axmacher, N. (2021). The memory trace of a stressful episode. Current Biology, 31(23), 5204-5213.

Boschen, M. J., Neumann, D. L., & Waters, A. M. (2009). Relapse of successfully treated anxiety and fear: theoretical issues and recommendations for clinical practice. Australian & New Zealand Journal of Psychiatry, 43(2), 89-100.

Bouton, M. E. (2002). Context, ambiguity, and unlearning: sources of relapse after behavioral extinction. Biological psychiatry, 52(10), 976-986.

Cavanna, A. E., & Trimble, M. R. (2006). The precuneus: A review of its functional anatomy and behavioural correlates. Brain: A Journal of Neurology, 129(Pt 3), 564-583. https://doi.org/10.1093/brain/awl004

Chaby, L. E., Karavidha, K., Lisieski, M. J., Perrine, S. A., & Liberzon, I. (2019). Cognitive flexibility training improves extinction retention memory and enhances cortical dopamine with and without traumatic stress exposure. Frontiers in Behavioral Neuroscience, 13, 24.

Craske, M. G., Treanor, M., Conway, C. C., Zbozinek, T., & Vervliet, B. (2014). Maximizing exposure therapy: An inhibitory learning approach. Behaviour research and therapy, 58, 10-23.

Cox, R. W., & Hyde, J. S. (1997). Software tools for analysis and visualization of fMRI data. NMR in Biomedicine, 10(4–5), 171–178. https://doi.org/10.1002/(SICI)1099-1492(199706/08)10:4/5<171::AID-NBM453>3.0.CO;2-L

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis: I. Segmentation and surface reconstruction. NeuroImage, 9(2), 179–194. https://doi.org/10.1006/nimg.1998.0395

Dosenbach, N. U. F., Fair, D. A., Miezin, F. M., Cohen, A. L., Wenger, K. K., Dosenbach, R. A. T., Fox, M. D., Snyder, A. Z., Vincent, J. L., Raichle, M. E., Schlaggar, B. L., & Petersen, S. E. (2007). Distinct brain networks for adaptive and stable task control in humans. Proceedings of the National Academy of Sciences of the United States of America, 104(26), 11073-11078. https://doi.org/10.1073/pnas.0704320104

Dunsmoor, J. E., Niv, Y., Daw, N., & Phelps, E. A. (2015). Rethinking Extinction. Neuron, 88(1), 47-63. https://doi.org/10.1016/j.neuron.2015.09.028

Esteban, O., Markiewicz, C., Blair, R. W., Moodie, C., Isik, A. I., Erramuzpe Aliaga, A., Kent, J., et al. (2018). fMRIPrep: A robust preprocessing pipeline for functional MRI. Nature Methods. https://doi.org/10.1038/s41592-018-0235-4

Exton-McGuinness, M. T., Lee, J. L., & Reichelt, A. C. (2015). Updating memories—the role of prediction errors in memory reconsolidation. Behavioural brain research, 278, 375-384.

Fonov, V. S., Evans, A. C., McKinstry, R. C., Almli, C. R., & Collins, D. L. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. NeuroImage, 47(Supplement 1), S102. https://doi.org/10.1016/S1053-8119(09)70884-5

Graner, J. L., Stjepanović, D., LaBar, K. S., & Dunsmoor, J. E. (2020). Aversive value generalization in human avoidance decision-making. Nature Communications, 11(1), 4839. https://doi.org/10.1038/s41467-020-18728-7

Greco, J. A., & Liberzon, I. (2016). Neuroimaging of fear-associated learning. Neuropsychopharmacology, 41(1), 320-334.

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. NeuroImage, 48(1), 63–72. https://doi.org/10.1016/j.neuroimage.2009.06.060

Gu, X., Wu, Y. J., Zhang, Z., Zhu, J. J., Wu, X. R., Wang, Q., ... & Li, W. G. (2022). Dynamic tripartite construct of interregional engram circuits underlies forgetting of extinction memory. Molecular Psychiatry, 27(10), 4077-4091.

Fullana, M. A., Harrison, B. J., 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. Molecular psychiatry, 21(4), 500-508.

Heinen, R., Bierbrauer, A., Wolf, O. T., & Axmacher, N. (2024). Representational formats of human memory traces. Brain Structure and Function, 229(3), 513-529.

Hennings, A. C., McClay, M., Drew, M. R., Lewis-Peacock, J. A., & Dunsmoor, J. E. (2022). Neural reinstatement reveals divided organization of fear and extinction memories in the human brain. Current Biology, 32(2), 304-314.

Huntenburg, J. M. (2014). Evaluating nonlinear coregistration of BOLD EPI and T1w images (Master’s thesis, Freie Universität, Berlin). http://hdl.handle.net/11858/00-001M-0000-002B-1CB5-A

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. NeuroImage, 17(2), 825–841. https://doi.org/10.1006/nimg.2002.1132

Josselyn, S. A., Köhler, S., & Frankland, P. W. (2015). Finding the engram. Nature Reviews Neuroscience, 16(9), 521-534.

Klein, A., Ghosh, S. S., Bao, F. S., Giard, J., Häme, Y., Stavsky, E., Lee, N., et al. (2017). Mindboggling morphometry of human brains. PLOS Computational Biology, 13(2), e1005350. https://doi.org/10.1371/journal.pcbi.1005350

Kobelt, M., Waldhauser, G. T., Rupietta, A., Heinen, R., Rau, E. M. B., Kessler, H., & Axmacher, N. (2024). The memory trace of an intrusive trauma-analog episode. Current Biology, 34(8), 1657-1669.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. Frontiers in systems neuroscience, 2, 249.

LaBar, K. S., & Phelps, E. A. (2005). Reinstatement of conditioned fear in humans is context dependent and impaired in amnesia. Behavioral neuroscience, 119(3), 677.

LaRocque, K. F., Smith, M. E., Carr, V. A., Witthoft, N., Grill-Spec

Lebois, L. A., Seligowski, A. V., Wolff, J. D., Hill, S. B., & Ressler, K. J. (2019). Augmentation of extinction and inhibitory learning in anxiety and trauma-related disorders. Annual review of clinical psychology, 15(1), 257-284.

Lee, J. L., Nader, K., & Schiller, D. (2017). An update on memory reconsolidation updating. Trends in cognitive sciences, 21(7), 531-545.

Lissek, S., & Tegenthoff, M. (2024). Dissimilarities of neural representations of extinction trials are associated with extinction learning performance and renewal level. Frontiers in Behavioral Neuroscience, 18, 1307825.

Liu, J. F., Yang, C., Deng, J. H., Yan, W., Wang, H. M., Luo, Y. X., ... & Lu, L. (2015). Role of hippocampal β-adrenergic and glucocorticoid receptors in the novelty-induced enhancement of fear extinction. Journal of Neuroscience, 35(21), 8308-8321.

Liu, Y., Ye, S., Li, X. N., & Li, W. G. (2024). Memory trace for fear extinction: fragile yet reinforceable. Neuroscience Bulletin, 40(6), 777-794.

Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. Behavior research methods, 49, 1494-1502.

Maren, S., Phan, K. L., & Liberzon, I. (2013). The contextual brain: Implications for fear conditioning, extinction and psychopathology. Nature Reviews. Neuroscience, 14(6), 417-428. https://doi.org/10.1038/nrn3492

Milad, M. R., Orr, S. P., Pitman, R. K., & Rauch, S. L. (2005). Context modulation of memory for fear extinction in humans. Psychophysiology, 42(4), 456-464.

Milad, M. R., & Quirk, G. J. (2012). Fear extinction as a model for translational neuroscience: ten years of progress. Annual review of psychology, 63(1), 129-151.

Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. Neuroimage, 59(3), 2636-2643.

Navarro-Sánchez, M., Gil-Miravet, I., Montero-Caballero, D., Castillo-Gómez, E., Gundlach, A. L., & Olucha-Bordonau, F. E. (2024). Some key parameters in contextual fear conditioning and extinction in adult rats. Behavioural Brain Research, 462, 114874.

Neumann, D. L. (2006). The effects of physical context changes and multiple extinction contexts on two forms of renewal in a conditioned suppression task with humans. Learning and Motivation, 37(2), 149-175.

Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. NeuroImage, 84(Supplement C), 320–341. https://doi.org/10.1016/j.neuroimage.2013.08.048

Ramanathan, K. R., Jin, J., Giustino, T. F., Payne, M. R., & Maren, S. (2018). Prefrontal projections to the thalamic nucleus reuniens mediate fear extinction. Nature communications, 9(1), 4527.

Redondo, R. L., Kim, J., Arons, A. L., Ramirez, S., Liu, X., & Tonegawa, S. (2014). Bidirectional switch of the valence associated with a hippocampal contextual memory engram. Nature, 513(7518), 426-430.

Rissman, J., Gazzaley, A., & D'Esposito, M. (2004). Measuring functional connectivity during distinct stages of a cognitive task. Neuroimage, 23(2), 752-763.

Rissman, J., & Wagner, A. D. (2012). Distributed representations in memory: insights from functional brain imaging. Annual review of psychology, 63(1), 101-128.

Santini, E., Quirk, G. J., & Porter, J. T. (2008). Fear conditioning and extinction differentially modify the intrinsic excitability of infralimbic neurons. Journal of Neuroscience, 28(15), 4028-4036.

Schiller, D., Levy, I., Niv, Y., LeDoux, J. E., & Phelps, E. A. (2008). From fear to safety and back: reversal of fear in the human brain. Journal of Neuroscience, 28(45), 11517-11525.

Schiller, D., & Delgado, M. R. (2010). Overlapping neural systems mediating extinction, reversal and regulation of fear. Trends in cognitive sciences, 14(6), 268-276.

Sommer, V. R., Mount, L., Weigelt, S., Werkle-Bergner, M., & Sander, M. C. (2022). Spectral pattern similarity analysis: Tutorial and application in developmental cognitive neuroscience. Developmental cognitive neuroscience, 54, 101071.

Sperl, M. F., Wroblewski, A., Mueller, M., Straube, B., & Mueller, E. M. (2021). Learning dynamics of electrophysiological brain signals during human fear conditioning. NeuroImage, 226, 117569.

Stevens, F. L., Hurley, R. A., & Taber, K. H. (2011). Anterior cingulate cortex: Unique role in cognition and emotion. The Journal of Neuropsychiatry and Clinical Neurosciences, 23(2), 121-125. https://doi.org/10.1176/jnp.23.2.jnp121

Szeska, C., Richter, J., Wendt, J., Weymar, M., & Hamm, A. O. (2020). Promoting long-term inhibition of human fear responses by non-invasive transcutaneous vagus nerve stimulation during extinction training. Scientific reports, 10(1), 1529.

Treiber, J. M., White, N. S., Steed, T. C., Bartsch, H., Holland, D., Farid, N., McDonald, C. R., Carter, B. S., Dale, A. M., & Chen, C. C. (2016). Characterization and correction of geometric distortions in 814 diffusion weighted images. PLOS ONE, 11(3), e0152472. https://doi.org/10.1371/journal.pone.0152472

Turner, B. O., Mumford, J. A., Poldrack, R. A., & Ashby, F. G. (2012). Spatiotemporal activity estimation for multivoxel pattern analysis with rapid event-related designs. NeuroImage, 62(3), 1429-1438.

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4ITK: Improved N3 bias correction. IEEE Transactions on Medical Imaging, 29(6), 1310–1320. https://doi.org/10.1109/TMI.2010.2046908

Vansteenwegen, D., Hermans, D., Vervliet, B., Francken, G., Beckers, T., Baeyens, F., & Eelen, P. (2005). Return of fear in a human differential conditioning paradigm caused by a return to the original acquistion context. Behaviour research and therapy, 43(3), 323-336.

Vervliet, B., Craske, M. G., & Hermans, D. (2013). Fear extinction and relapse: state of the art. Annual review of clinical psychology, 9(1), 215-248.

Visser, R. M., Scholte, H. S., & Kindt, M. (2011). Associative learning increases trial-by-trial similarity of BOLD-MRI patterns. Journal of Neuroscience, 31(33), 12021-12028.

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., 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. https://doi.org/10.1016/j.psyneuen.2015.01.021

Wang, S., Peterson, D. J., Gatenby, J. C., Li, W., Grabowski, T. J., & Madhyastha, T. M. (2017). Evaluation of field map and nonlinear registration methods for correction of susceptibility artifacts in diffusion MRI. Frontiers in Neuroinformatics, 11. https://doi.org/10.3389/fninf.2017.00017

Wisniewski, D., Braem, S., González-García, C., De Houwer, J., & Brass, M. (2023). Effects of Experiencing CS–US Pairings on Instructed Fear Reversal. Journal of Neuroscience, 43(30), 5546-5558.

Xu, W., & Südhof, T. C. (2013). A neural circuit for memory specificity and generalization. Science, 339(6125), 1290-1295.

Xin, L., Ying, M., Qi, W., & Yi, L. (2024). Fear Reversal Learning: A New Method of Fear Regulation. Journal of Psychological Science, 47(2), 494.

Xue, G., Dong, Q., Chen, C., Lu, Z., Mumford, J. A., & Poldrack, R. A. (2010). Greater neural pattern similarity across repetitions is associated with better memory. Science, 330(6000), 97-101.

Xue, G., Dong, Q., Chen, C., Lu, Z. L., Mumford, J. A., & Poldrack, R. A. (2013). Complementary role of frontoparietal activity and cortical pattern similarity in successful episodic memory encoding. Cerebral Cortex, 23(7), 1562-1571.

Xue, G. (2018). The neural representations underlying human episodic memory. Trends in cognitive sciences, 22(6), 544-561.

Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Transactions on Medical Imaging, 20(1), 45–57. https://doi.org/10.1109/42.906424

Zheng, L., Gao, Z., Xiao, X., Ye, Z., Chen, C., & Xue, G. (2018). Reduced fidelity of neural representation underlies episodic memory decline in normal aging. Cerebral Cortex, 28(7), 2283-2296.