**RESULTS**

Each subject encoded trial-unique pictures of animals and tools before (baseline, i.e., pre-conditioning), during, and after (i.e., extinction) fear conditioning. One semantic category (animals or tools, counterbalanced) served as CS+ and co-terminated with an electrical shock on 50% of trials during fear conditioning, while the other category (tools or animals, respectively) were never paired with shock (CS-). Shocks were only delivered during fear conditioning. As detailed elsewhere (Hennings et al. 2020), the extinction temporal context was further distinguished from the conditioning context by incorporating task-irrelevant pictures of scenes during the intertrial interval between each CS trial. Participants returned 24-hours later for a surprise recognition memory test composed of all CSs encoded during each phase, as well as novel lures. Participants were informed that no shocks would be delivered during the memory test. The analysis here focuses on the overlap of multi-voxel activity patterns of items from encoding to retrieval (i.e., encoding retrieval similarity, ERS), irrespective of memory performance.



**Figure 1. Divided organization of opposing long-term fear and extinction memory in the human brain**. **A**. **Schematic overview.** Neurobiological research reveals competing experiences of fear and extinction are maintained as distinct memory traces in the brain. This divided organization is adaptive for mitigating fear overgeneralization to stimuli that are known to be safe, while also maintaining threat associations given a countervailing experience of safety. Whether opposing memories of fear and extinction can be isolated in the human brain remains unclear. **B. Simplified circuits diagrams** of fear and extinction memory retrieval, highlighting the interactions between the MTL and mPFC. Human homologues of neural structures in rodents are given in parentheses. **C.** **Overview of the associative learning task on Day 1.** Semantic categories of images served as the CS+/-, each trial was a unique category exemplar that did not repeat. For example, only one image of a cow was shown on Day 1. During fear conditioning, 50% of the CS+ co-terminated with a mild electric shock (US). During extinction learning, the normal ITI was replaced by a stream of natural scene images to build the extinction context tag. **D. Overview of the encoding-retrieval similarity analysis.** 24-hrs after associative learning, participants were placed back into the scanner and completed a surprise recognition memory test for the items encoded on Day 1. Each trial during encoding and retrieval generates a unique pattern of activity. To test for neural reinstatement, the encoding and retrieval patterns elicited by a single image are correlated within a given ROI.

**Behavioral results**

*Associative learning*. We have previously reported behavioral results from the associative learning task on Day 1, and replicate them here for clarity. The success of fear conditioning and extinction learning was assayed by autonomic arousal (skin conductance responses; SCR) and explicit shock expectancy (Yes/No 2-alternative forced choice). Analyses focused on differential responding (i.e. CS+ > CS- differences) in SCR and shock expectancy from each phase. During conditioning, both healthy adults and individuals with PTSS exhibited significant CS+ > CS- responses for both SCR (Healthy: t(23) = 4.22, P = 3.25e-4; PTSS: t(23) = 3.17, P = 4.31e-3) and shock expectancy (Healthy: t(23) = 14.3, P = 6.16e-13; PTSS: t(23) = 7.62, P = 9.89e-8). The success of extinction learning was assessed by comparing differential responses from conditioning to the second half of extinction. Both groups displayed a significant reduction in differential SCR (Healthy: t(21) = -2.60, P = 0.017; PTSS: t(21) = -2.86, P = 9.34e-3) and shock expectancy (Healthy: t(23) = -4.33, P = 2.46e-4; PTSS: t(23) = -3.66, P = 1.29e-3). Critical to the present report, there were no significant differences in behavioral responses between healthy participants and participants with PTSS during either conditioning (SCR: t(46) = 0.63, P = 0.53; expectancy: t(46) = 1.23, P = 0.22) or the second half of extinction (SCR: t(42) = 0.49, P = 0.63; expectancy: t(46) = 0.69, P = 0.50). Together these results demonstrate successful fear conditioning and subsequent extinction across both groups.

Chart

Description automatically generated

**Figure 2. Dissociable reinstatement of emotional memories across the mPFC. A. ERS in *a priori* ROIs.** CS+ - CS- ERS in *a priori* dACC and vmPFC masks taken from previous literature.Error bars correspond to the 95% confidence interval of the CS+/- difference. Asterisks on single bars denote significance of CS+ - CS- for that phase/group. Pairwise asterisks indicate significant difference between ROIs. \*\*\*P < 0.001, \*\*P < 0.01, \*P < 0.05 FDR corrected. *Top.* Healthy adults exhibited a double dissociation of emotional reinstatement in the mPFC, such that reinstatement for items encoded during conditioning was higher in the dACC, and extinction reinstatement was higher in the vmPFC. *Bottom*. In PTSS, the dACC displayed significant emotional reinstatement of items encoded during both conditioning and extinction, representing a misallocation of extinction memories. **B. Whole brain ERS searchlight.** Results confirm *a priori* results and reveal other regions engaged in selective emotional reinstatement. A medial and lateral view of the left hemisphere is shown, and results were qualitatively similar across hemispheres. The heatmaps show average effect of CS+ - CS- reinstatement, each map was threshold at P < 0.001 one-sided for CS+ > CS- with a cluster-wise threshold (FWE) of P < 0.05.

**Selective and dissociable encoding-retrieval similarity of fear and extinction in the mPFC.**

The voxel-wise pattern of activity elicited by each CS item encoded at each encoding phase (pre-conditioning, fear conditioning, or extinction) was correlated with the pattern of activity elicited by that same CS item at retrieval. For details on the full statistical approach see Methods. To evaluate selective item-level ERS to the CS+ versus CS-, correlations were fisher z transformed, and the average correlation of CS- trials was subtracted from the average correlation of the CS+ trials from the same encoding context. In this way we control for general item-level reinstatement to focus on selective retrieval similarity for emotionally relevant stimuli in the mPFC. The multivariate analysis of encoding-retrieval similarity focused on two distinct subregions of the mPFC that rodent research shows to play dissociable roles in the encoding, storage, and retrieval of fear and extinction memory, the dACC and vmPFC. Regions were defined *a priori* by drawing spheres around peak activations from pervious literature, see Methods for more details.

In healthy adults, the dACC did not exhibit selective reinstatement of CS+ versus CS- items encoded prior to fear conditioning. This confirms that ERS was not different between CSs which had not yet acquired emotional value at the time of encoding and provides a baseline manipulation check that this region does not indiscriminately reinstate CS+ items regardless of when they were encoded. For items encoded during fear conditioning, the dACC exhibited selective reinstatement for CS+ (CS diff. = 0.22, [0.16, 0.28], PFDR = 4.62e-12). This finding accords with rodent models that show the prelimbic cortex is involved in both the learning and retrieval of long-term fear memories. For items encoded during extinction, the dACC did not show selective reinstatement for CS+. The difference in selective CS+ reinstatement in the dACC between items encoded in the fear and extinction context was significant (Phase diff. = 0.21, [0.12, 0.29], P = 1.57e-6).

In the vmPFC, there was likewise no selectivity in ERS between CS+ and CS- for items encoded prior to fear conditioning in healthy adults. There was selective reinstatement in the vmPFC for CS+ items encoded in both the conditioning (CS diff. = 0.074, [0.013, 0.134], PFDR = 0.033) and extinction (CS diff. = 0.113, [0.053, 0.173], PFDR = 9.20e-4) contexts. Notably, there was significantly more selective CS+ reinstatement for items encoded during conditioning in the dACC versus the vmPFC (ROI diff. = 0.149, [0.064, 0.234], PFDR = 0.002). The converse was also true; that is, there was significantly more selective CS+ reinstatement for items encoded during extinction in the vmPFC versus the dACC (ROI diff. = 0.099, [0.014, 0.184], PFDR = 0.031). Thus, in healthy adults, discrete regions of the mPFC exhibited a double dissociation in selective reinstatement based on the temporal context in which the CS+ memory was formed.

Individuals with PTSS also exhibited reinstatement of CS+ items in the dACC for items encoded during conditioning (CS diff. = 0.171, [0.111, 0.231], PFDR = 1.53e-7), indicating a pattern of selective fear memory reinstatement consistent with healthy adults. Similar to the healthy adult group, there was also stronger reinstatement of fear memories in the dACC than in the vmPFC (ROI diff. = 0.121, [0.036, 0.206], PFDR = 0.011). Unlike the healthy adult group, however, the PTSS group showed selective reinstatement in the dACC for CS+ items encoded in the extinction context as well (CS diff. = 0.103, [0.043, 0.164], PFDR = 0.002). That information encoded in the extinction context was reinstated in the same region involved in the formation and retrieval of fear memories suggests a misallocation of extinction memories. Also in contrast to the healthy adult group, there was no evidence of selective reinstatement for CS+ items encoded during either conditioning or extinction in the vmPFC. The vmPFC did exhibit an unexpected selectivity for CS- items encoded prior to fear conditioning in the PTSS group (CS diff. = -0.079, [-0.139, -0.019], PFDR = 0.024), which we speculate upon in the discussion (OR maybe in a supplementary discussion?).

**Whole brain analysis reveals widespread selective cortical reinstatement.**

Research utilizing rodent models are often limited in terms of the number of neural structures that can be observed in a given experiment. With fMRI, we have simultaneous access to the whole brain. Given the results we obtained using our *a priori* ROIs, we conducted an exploratory whole brain searchlight analysis seeking selective emotional reinstatement beyond our *a priori* ROIs. In brief, this analysis iterates over all possible spheres of voxels and calculates the local encoding-retrieval similarity at each location (see Methods). The resulting ERS maps undergo family-wise error correction to identify significant clusters of voxels that exhibit CS+ > CS- reinstatement. For each group, searchlight maps were taken corresponding to the CS+ - CS- difference in reinstatement the emotional encoding contexts (conditioning and extinction) (**Fig 2B**).

In healthy adults, the searchlight analysis confirmed the results of our *a priori* analysisand revealed other regions which exhibited significant selective emotional reinstatement (See **Supplementary Table 1** for full list of cluster locations). As expected, for items encoded during conditioning there were bilateral clusters which qualitatively overlapped with our *a priori* dACC and vmPFC masks. In addition, there were significant clusters in the anterior insula, a region that has previously been identified as involved in the acquisition and recall of conditioned fear in humans (Fullana et al., 2016). For extinction, the largest cluster qualitatively overlapped with our vmPFC mask (medial frontal gyrus). For both phases, several other frontal and posterior regions of cortex were identified as well, which suggest that the neural representations of fear and extinction are distributed across various regions of cortex. Several regions (e.g. medial frontal gyrus and precuneus) exhibited selective emotional reinstatement for fear and extinction, which may indicate that in healthy adults these regions play a more general role in emotional memory processes regardless of memory valence.

Individuals with PTSS displayed similar whole-brain searchlight results for conditioning as healthy adults, with large clusters corresponding to the dACC, bilateral insula, as well as other cortical regions. For extinction, we observed significant clusters in the cuneus, as well as bilateral insula. The insula clusters are of particular interest, as this is another example of individuals with PTSS misallocating extinction memories to a region that normally codes for threat memories. We did not observe significant extinction in the dACC in the searchlight as we did in our *a priori* ROI, which is most likely due to the difference in size between our a priori mask (10mm radius) and the size of the searchlight spheres (6mm radius). Overall, the searchlight analysis confirmed the pattern of results obtained in our *a priori* analysis, as well as demonstrating that in humans, emotional memory retrieval elicits significant reinstatement beyond the mPFC to other cortical regions.

**Contextual specificity of encoding-retrieval similarity in the medial temporal lobe**

A surprising null result from the whole brain searchlight analysis was the lack of subcortical clusters exhibiting selective emotional reinstatement for either conditioning or extinction. Previous work has shown that the amygdala and hippocampus are core components of fear and extinction neurocircuitry involved in the acquisition and retrieval of both fear and extinction related memories. The hippocampus in particular exerts contextual control over memory retrieval. Emerging neurobiological models in rodents indicate that different subfields along the long-axis of the hippocampus serve discrete functions in the course of conditioning and extinction. Human neuroimaging also shows functional specializations for these subfields in memory and affective processes. Using subject-specific anatomical segmentations, we probed ERS along the long-axis of the hippocampus in three bi-lateral subfields: head (anterior), body, and tail (posterior). The amygdala was similarly segmented into two bilateral ROIs: the basolateral amygdala (BLA) and the central nucleus of the amygdala (CeM). These amygdalar subfields also known to have functional specialization in conditioning and extinction processes.

*Hippocampus.* As with our analysis of ERS in the mPFC, we examined whether subfields of the hippocampus exhibited selective emotional reinstatement. Selective (CS+ - CS-) reinstatement was not observed for any encoding context in any subfield, in either group (all PFDR > 0.45). Even though there was no evidence for CS specific reinstatement, a linear-mixed effects model revealed a significant three-way interaction of *encoding context \* subfield \* group* (X2(4­) = 12.8, P = 0.012; see Methods for full model specification). The significance of this term suggests that subfields of the hippocampus may be sensitive to encoding context in general, but not CS type. As such, we probed the specificity of encoding context specific reinstatement, averaging across CS+/-.

In both groups, the posterior hippocampus exhibited significant selectively in ERS for the fear conditioning context. In healthy adults, ERS of items encoding during conditioning was significantly greater than items encoded during extinction (phase diff = 4.36e-2, [1.40e-2, 7.32e-2], PFDR = 0.019). In PTSS, ERS of items encoding during conditioning was greater than items encoded during pre-conditioning (phase diff. = 4.41e-2, [1.45e-2, 7.37e-2], PFDR = 0.019) and extinction (phase diff. = 4.33e-2, [1.36e-2, 7.29e-2], PFDR = 0.019). The body of the hippocampus did not exhibit any encoding context specific reinstatement. In contrast to the posterior, the anterior hippocampus displayed significant ERS selectively for the extinction context, however this was only observed in healthy adults. Specifically, ERS of items encoding during extinction was greater than items encoding during fear conditioning in healthy adults (phase diff. = 0.065, [0.035, 0.094], PFDR = 3.36e-4).

These results suggest a gradient of functional specialization along the long axis of the hippocampus, with the posterior showing preference for the fear conditioning context, and the anterior for the extinction context. We tested this dissociation directly by contrasting the amount of reinstatement for each context between the posterior and anterior hippocampus (**Fig. 3**). There was a significant double dissociation in ERS between the subfields, but only for healthy adults. Specifically, ERS for items encoded in the conditioning context was greater in the posterior compared to the anterior hippocampus (ROI diff. = 0.033, [0.003, 0.063], PFDR = 0.038), and ERS for items encoded in the extinction context was greater in the anterior hippocampus compared to the posterior (ROI diff. = 0.075, [0.046, 0.105], PFDR = 2.51e-6). As in healthy adults, in PTSS the posterior hippocampus showed more conditioning context ERS compared to the anterior hippocampus (ROI diff. = 0.034, [0.004, 0.063], PFDR = 0.038). However, PTSS did not show any difference between the posterior and anterior hippocampus in terms of extinction context ERS (ROI diff. = -0.004, [-0.034, 0.026], PFDR = 0.80). The lack of extinction reinstatement in the anterior hippocampus further supports the idea that the neural organization of safety memories is dysregulated in PTSS as compared to healthy adults.



**Figure 3. Differential ERS for emotional contexts along the long axis of the hippocampus.** ERS was averaged across CS+/- in each subfield of the hippocampus.Error bars correspond to the 95% confidence interval of the marginal means. Significant phase-specific ERS was observed in the posterior and anterior subfields, but not the body of the hippocampus (data not shown). Pairwise asterisks indicate significant difference between subfields. \*\*\*P < 0.001, \*P < 0.05 FDR corrected. *Top.* Healthy adults exhibited a double dissociation of reinstatement in the hippocampus, such that reinstatement for items encoded during conditioning was higher in the posterior, and extinction reinstatement was higher in the anterior. *Bottom*. In PTSS, the posterior subfield exhibited more reinstatement of conditioning items than the anterior hippocampus.

*Amygdala*. We also examined whether subfields of the amygdala exhibited selective (CS+ - CS) emotional reinstatement, however none was observed for any encoding context in any subfield, in either group (all PFDR > 0.64). In addition, we did not observe any significant main effects or interactions in a linear mixed-effects model, and thus did not preform any other follow-up tests. Reasons for the lack of significant ERS in the amygdala are explored in the discussion.

**MTL activity at retrieval predicts dissociable reinstatement in the mPFC**

*Univariate activity.* Our *a priori* analysis of ERS in the mPFC showed that healthy adults exhibited a double dissociation of emotional reinstatement, with more conditioning-specific CS+ reinstatement in the dACC and more extinction-specific CS+ reinstatement in the vmPFC. What determines which area of the mPFC a particular item is reinstated in? As we have described, previous work has shown that the hippocampus and amygdala are core components of the neural circuits which mediate emotional memory retrieval. We hypothesized that neural activity in these subcortical regions would be predictive of where in the mPFC an item is preferentially reinstated. On a trial-by-trial basis we tested whether univariate activity in different subcortical regions predicted the difference in ERS between our two mPFC regions (vmPFC – dACC). We restricted our analysis to items encoded during conditioning and extinction, and tested in turn neural activity from the posterior, body and anterior hippocampus, amygdala BLA and CeM. In this analysis, if a predictor has a significant positive slope, this indicates it predicts a bias towards the vmPFC, while a significant negative slope indicates the predictor predicts a bias towards the dACC.

We found that all subfields of both the hippocampus and amygdala were significant negative predictors, which indicates that increases in subcortical univariate activity bias reinstatement to the dACC (posterior HC: X2(1) = 54.7, P = 1.38e-13, slope = -1.8e-3; HC body: X2(1) = 68.2, P = 1.48e-16, slope = -2.49e-3; anterior HC: X2(1) = 46.8, P = 8.00e-12, slope = -1.7e-3; BLA: X2(1) = 26.7, P = 2.39e-7, slope = -1.45e-3; CeM: X2(1) = 19.5, P = 1.01e-5, slope = -6.90e-4). Additionally, we observed several interactions with hippocampal subfields. For the posterior hippocampus, there was a significant *posterior HC*\**CS* *condition* (Chisq (1) = 11.2, P = 8.3e-4), such that the slope of posterior hippocampal activity was significantly more negative for CS+ compared to CS- (CS slope diff. = -1.53e-3, [-2.43e-3, -6.34e-4], P = 8.24e-4). In the body of the hippocampus, there was a significant *HC body*\**CS condition*\**encode phase* interaction (Chisq (1) = 5.46, P = 0.019). Post-hoc contrasts revealed that for items encoded during conditioning, the slope for the CS+ was significantly more negative than the CS- (CS slope diff. = -2.68e-3, [-4.21e-3, -1.15e-3], PFDR = 1.17e-3), while there was no difference in the slopes for extinction (CS slope diff. = -8.68e-5, [-1.63e-3, 1.46e-3], PFDR = 0.91). There were no significant interactions in the anterior hippocampus, BLA, or CeM.

In sum, subcortical univariate activity predicated more reinstatement in the dACC. This prediction was stronger for all CS+ in the posterior hippocampus compared to CS-, and was selective for conditioning CS+ items in the body of the hippocampus.

*Subcortical ERS*. Next, we conducted a similar set of analyses, in which trial-by-trial ERS was used to predict the vmPFC – dACC difference in reinstatement instead of univariate activity. Even though univariate activity across the hippocampus and amygdala all predicted a bias in reinstatement towards the dACC, this is not automatically the hypothesis for what ERS will predict. The information represented in a pattern of activity, captured by ERS, is a different signal from the total average activation across that pattern.

Unlike univariate activity, we found both significant positive and negative predictors of the mPFC ERS bias (**Fig 4**). Trial-by-trial reinstatement in both the posterior hippocampus and CeM were both significant negative predictors (posterior HC: X2(1) = 4.64, P = 0.031, slope = -0.060; CeM: X2(1) = 8.49, P = 0.004, slope = -0.065), indicating that more reinstatement in these regions is associated with a bias in mPFC reinstatement towards the dACC. In contrast, the anterior hippocampus was a positive predictor (X2(1) = 11.1, P = 8.51e-4, slope = 0.091), meaning that more reinstatement in this region is associated with a bias in mPFC reinstatement towards the vmPFC. ERS in the body of the hippocampus and the amygdala BLA were not significant predictors. There were no significant interactions in the models for the CeM, posterior, or anterior hippocampus; indicating the observed relationships between subcortical and mPFC ERS are general, and not dependent on encoding context, CS type, or group.

Chart

Description automatically generated

**Figure 4. Predicting the bias in mPFC reinstatement. A. Descriptive overview of the mPFC difference in ERS.** A kernel density plot shows the overall distribution of vmPFC – dACC ERS for CS+/- from conditioning and extinction, collapsed across group. **B. Subcortical ERS predicts mPFC difference.** For each of the subcortical ROIs, local ERS for each trial was used to predict the overall difference of ERS in the mPFC. A positive slope indicates local ERS predicts a bias in reinstatement towards the vmPFC, while a negative slope predicts a bias towards the dACC. The terms shown reflect the predictiveness of ERS in each subcortical ROI in general, across all levels of phase, CS type, and group. Points and error bars correspond to the estimate and 95% confidence interval of the slope of subcortical ERS in each model. \*\*\*P < 0.001, \*\*P < 0.01, \*P < 0.05.

*Separable influence of the anterior hippocampus.* We found that univariate and multivariate signals from the anterior hippocampus predict opposite biases in mPFC reinstatement during memory retrieval. Anterior hippocampal univariate activity predicted a bias towards the dACC, while ERS in the same region predicted more reinstatement towards vmPFC. A region which exhibits bi-directional control could be crucial for the proper regulation of fear and extinction in humans. As such, we tested whether these effects were separable in a single model: using anterior hippocampal univariate activity and ERS as simultaneous predictors of the mPFC difference in ERS. Indeed, in this model both predictors retained their significance and sign (univariate: X2(1) = 42.5, P = 7.1e-11, slope = -1.65e-3; ERS: X2(1) = 5.56, P = 0.018, slope = 0.067). Again, there were no significant interactions with either predictor. These results suggest that in the anterior hippocampus, more total activity predicts a bias in reinstatement to the dACC, while more local reinstatement predicts a bias in reinstatement to the vmPFC.

**Reinstated extinction context predicts bias in reinstatement to vmPFC.**

For reasons detailed in Hennings et al., (2020), trial irrelevant “context tags” were inserted between each trial during extinction learning on day 1. These context tags serve as a proxy for the context in which they are encoding in, and have been used to track the reinstatement of the encoding context during memory retrieval (Bornstein & Norman, 2017; Gershman et al., 2013; Manning et al., 2016). During the fear renewal test, we demonstrated the context tag encoded during extinction learning served as proxy for the associative memory of extinction: in healthy adults the degree of extinction context reinstatement predicted the behavioral expression of extinction and neural activity in the vmPFC. Here, we explored whether evidence for the reinstatement of the extinction context during the recognition memory test predicted the bias in mPFC reinstatement. Based on our previous work, we predicted there was a relationship, reinstated extinction context would predict a bias in mPFC reinstatement to the vmPFC.

Similar to our analyses relating subcortical univariate and ERS mPFC reinstatement, trial-by-trial levels of extinction context evidence were used to predict the differential split in mPFC ERS. We again limited our analysis to conditioning and extinction trials as our encoding contexts of interest. Extinction context evidence was a significant positive predictor (X2(1) = 12.6, P = 3.92e-4, slope = 0.093). In addition, there was a trending *extinction context*\**CS condition* interaction (X2(1) = 3.66, P = 0.056). We explored this interaction by testing the marginal slope of *extinction context* in each CS, collapsed across phase and group. Reinstated extinction context was a significant positive predictor for CS+ (slope = 0.139, [0.069, 0.208], PFDR = 1.84e-4), but was not significant for CS- (slope = 0.048, [-0.021, 0.117], PFDR = 0.18). These results demonstrate neural evidence for the previously encoded extinction context is associated with a bias in mPFC ERS towards the vmPFC, selective to emotional CS+ items.

Diagram

Description automatically generated

**Figure 5. Reinstated extinction context predicts a bias in mPFC reinstatement towards the vmPFC.** During the ITIs of extinction learning on Day 1, participants are shown a stream of natural scene images. The next day, a MVPA classifier decodes evidence for scene images from the PPA, when no scene images are actually shown. Classifier evidence for scenes is thus a proxy for reinstatement of the extinction context. The classifier is trained on data from a perceptual localizer collected after the recognition memory test. Extinction context evidence predicted a bias in mPFC reinstatement towards the vmPFC, but only for CS+ images. This is consistent with a role for the vmPFC in extinction retrieval. Points and error bars correspond to the estimate and 95% confidence intervals of the marginal slopes for each CS type. \*\*\*P < 0.001.