**Title:**

Divided organization of fear and extinction memories in the human brain

Dissociable reinstatement of fear and extinction memories reveals the divided organization of emotional memories in the human brain

Dissociable neural reinstatement of fear and extinction memories in human mPFC

**Authors:**

Augustin C. Hennings, Mason McClay, Michael Drew, Jarrod A. Lewis-Peacock, Joseph E. Dunsmoor

**Abstract**

Neurobiological research in rodents has revealed that competing experiences of fear and extinction are stored as distinct memory traces in the brain. This divided organization is adaptive for mitigating overgeneralization of fear to related stimuli that are learned to be safe, while also maintaining threat associations for unsafe stimuli. Whether a similar dissociation exists in the human brain remains unclear. Here, we used a hybrid form of Pavlovian conditioning with an episodic memory component to identify multivariate patterns of fMRI activity associated with the formation and retrieval of fear versus extinction. In healthy adults, distinct regions of the medial PFC and hippocampus showed selective neural coding for fear and extinction memories. This dissociation was absent in participants with PTSD symptoms. The divided neural organization of fear and extinction may support flexible retrieval of context-appropriate emotional memories, while their disorganization may promote overgeneralization and increased fear relapse in affective disorders.

**Introduction**

Maintaining separate and competing memories of threat and safety is key to adaptive behavior. The inability to maintain memories of safety to overcome threat associations is characteristic of affective disorders such as posttraumatic stress disorder (PTSD) (Lissek and van Meurs, 2015). Neurobiological research using Pavlovian conditioning as a model shows that neural ensembles within and between dissociable regions organize the encoding, storage, and retrieval of fear (threat) and extinction (safety) memory (Etkin and Wager, 2007; Maren, 2001; Quirk and Mueller, 2008). This research confirms early theories—dating back to the time of Pavlov—that extinction is an active learning process that generates a secondary memory of safety that is stored in parallel to the memory of fear. In the rodent brain, these memory traces can be separated into discrete neural ensembles with distinct pathways between regions of the medial temporal lobe and subdivisions of the medial prefrontal cortex (mPFC) (Sierra-Mercado et al., 2011). Whether a similar neural organization exists in the human brain, whereby fear and extinction memories are segregated into separate neural regions, is unclear. Here, we use multivariate pattern analysis (MVPA) of functional magnetic resonance imaging (fMRI) data to isolate spatially distributed patterns of activity unique to the encoding and retrieval of fear versus extinction memories. We compare these neural signatures between healthy adults and individuals with PTSD symptoms, for which the ability to organize separable fear and extinction memories is presumably abnormal (Milad and Quirk, 2012).

Identifying quantifiable memory traces in the brain can be challenging: memory representations are widely distributed within and across discrete brain regions (Johnson et al., 2009; Polyn et al., 2005; Ritchey et al., 2013; Staresina et al., 2012; Staudigl et al., 2015), memories change over time (Gilmore et al., 2021; Sederberg et al., 2011) , and not all experiences induce a persistent change in the brain (Redondo and Morris, 2011). Fear conditioning is an ideal model to investigate the neural representations of memory, as it rapidly induces a stable and persistent associative memory with an objective behavioral correlate (LaBar et al., 1998). One of the most important discoveries in the neuroscience of associative learning has been the localization of neural circuits selective for the formation and retrieval of fear versus extinction memory. In the MTL sparse coding allows for fear and extinction to exist simultaneously in the same structures (Frankland et al., 2019; Josselyn et al., 2015), while a more stark division exists in the mPFC. The prelimbic cortex, homologous to the human dorsal anterior cingulate cortex (dACC), is activated during learning and retrieval of fear associations (Burgos-Robles et al., 2009; Sotres-Bayon et al., 2012), whereas the infralimbic cortex, homologous to the human ventromedial PFC (vmPFC), is a critical site of extinction memory formation and retrieval (Do-Monte et al., 2015; Klavir et al., 2017; Milad and Quirk, 2002). These areas interact dynamically with the amygdala and hippocampus to determine expression or suppression of conditioned fear (Marek et al., 2018; Senn et al., 2014).

Human neuroimaging has successfully translated evidence that the dACC is among the most consistently active regions during fear conditioning (Fullana et al., 2016a; Milad et al., 2007). Whether the human dACC is also a site of long-term storage and retrieval of fear memories is far less clear. Moreover, neuroimaging evidence of vmPFC involvement in extinctionis surprisingly scant. Indeed, meta-analyses show the vmPFC is *not* among a collection of regions active during extinction learning (Fullana et al., 2018). This inconsistency between animal neurophysiology and human neuroimaging has been a puzzle and limits the translational utility of advances in extinction research from rodents to humans.

A major hurdle to translating animal neurophysiology to human neuroimaging is a methodology to “tag” brain activity uniquely associated with competing memories of fear and extinction. In rodents, state-of-the-art advances in activity-dependent labeling can separate these memory traces by measuring the overlap in neuronal activity during acquisition and retrieval in collections of neurons, termed engrams (Frankland et al., 2019; Lacagnina et al., 2019). An analogous analytic approach in human neuroimaging involves correlating multivariate patterns of activity during memory encoding and retrieval. The match between patterns of activity in distributed voxels during encoding and retrieval (encoding-retrieval similarity, ERS) provides an index of memory fidelity, albeit not at the cellular level. This neuroimaging technique has been widely applied to the study of human memory (Johnson et al., 2009; Polyn et al., 2005; Ritchey et al., 2013; Staresina et al., 2012; Staudigl et al., 2015). Whether this technique can be leveraged to isolate separate memory traces of fear and extinction has not been tested, but we believe that it could provide a more precise index for how these competing memories are maintained in the human brain and altered in psychiatric disease.

Here, we use a hybrid conditioning/episodic memory design that incorporates trial-unique (i.e., non-repeating) semantic exemplars as conditioned stimuli (CS) during fear conditioning and extinction (Dunsmoor and Kroes, 2019). The next day subjects return to the MRI for a memory test composed of the unique CSs encoded during both conditioning and extinction. This hybrid conditioning design overcomes an inherent obstacle to the typical conditioning protocol. That is, typically the same CS (e.g., a colored shape) is repeated across all experimental phases; thus, it is only possible to measure retrieval of either the putative fear or extinction memory at test, not both. By incorporating trial-unique CSs, we can simultaneously “tag” and track specific episodes associated with either fear or extinction by correlating patterns of activity elicited by each CS during encoding and retrieval. In this way, we can quantify whether and how these competing memories are distinctly organized in separable patterns of activity in the same subject and in a single experiment. This modification allows us to adapt technical advances in neuroimaging of human episodic memory with the conceptual framework of functional labeling from rodent neurophysiology. This approach could provide a substantial translational leap forward by bridging advances in the neuroscience of fear and extinction memory from rodents to more precisely localize these memories in the human brain.

We predicted that the healthy adult brain organizes and maintains separable mnemonic representations based on the temporal context in which the memory was originally formed, i.e., fear conditioning or extinction learning. Such an organization is adaptive, as it prevents overwriting memory of threat given a countervailing experience of safety, and it helps ensure that the appropriate memory can be retrieved and expressed at the appropriate time. We directly compared neural signatures between healthy adults and individuals with post-traumatic stress symptoms (PTSS). We hypothesize that fear memory will be similarly represented in healthy adults and PTSS, but that the encoding and retrieval of extinction memory be different for these groups. Because of the maladaptive return of fear in PTSD, it could be that information encoded during extinction for this group is misallocated to regions involved in maintaining fear memories.

**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., 2020a), 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.** People maintain competing representations of danger and safety for closely related stimuli or situations. We can often retrieve the appropriate associative memory given the context. How the brain organizes these competing memories remains an important question, as disorganization between these competing representations may lead to maladaptive fear and anxiety in harmless situations. **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.

A picture containing text, clock

Description automatically generated

**Figure 2. Behavioral responses during fear conditioning and extinction.** Autonomic arousal and explicit shock expectancy from fear conditioning and late extinction (2nd half) on Day 1 are replicated from Hennings et al., 2020. Results show successful acquisition and extinction of differential (CS+ > CS-) responses for both SCR and shock expectancy for both groups. Critically, no group differences were observed in behavioral responses during associative learning. See text for statistical analyses. Error bars correspond to the 95% confidence interval of the mean.

**Behavioral results**

*Associative learning*. The success of fear conditioning and extinction learning was assessed by skin conductance responses (SCR) and trial-by-trial shock expectancy (Yes/No 2-alternative forced choice; see also Hennings et al., 2020). 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) = -2.6, 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). Importantly, 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 and equivalent fear conditioning and within-session extinction in both groups.

*Recognition memory.* Overall, participants had good performance on the recognition memory test (**Supplementary Figure 1**). The results replicated previous behavioral findings (Dunsmoor et al., 2015, 2018; Keller and Dunsmoor, 2020), in that memory was better for CS+ items compared to CS- from all phases, and overall higher for conditioning compared to other phases. For an analysis of corrected recognition memory performance, see Hennings et al., 2021 (In press). Here, we report an analysis of high confidence hit rates with the purpose of confirming no differences in behavioral performance between healthy adults and individuals with PTSS during the recognition memory test. A mixed-effects ANOVA of high confidence hit rates revealed no significant main effect of *group* (F1, 46 = 1.37, P = 0.25), and no significant higher order interactions between *group* and *CS type, encoding context,* or a three-way interaction(All Ps > 0.44). These results indicate that behavioral responding during the recognition memory was equivalent between 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. 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. These regions were defined *a priori* by drawing spheres around peak activations from previous literature (see Methods).

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.

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

To complement the results we obtained using our *a priori* ROIs of the dACC and vmPFC, we conducted an exploratory whole brain searchlight analysis seeking selective CS+ > CS- reinstatement across the brain. This analysis iterated over all possible spheres of voxels and calculated the local encoding-retrieval similarity at each location. The resulting ERS maps underwent 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 re-confirmed the *a priori* ROI analysisand also revealed additional brain regions exhibiting selective reinstatement of fear and extinction memory (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 dACC and vmPFC. In addition, there were significant clusters in the anterior insula, a region consistently identified in the fear acquisition and expression in human (Fullana et al., 2016b). For extinction, the largest cluster qualitatively overlapped with vmPFC (medial frontal gyrus). For both emotional phases, several other frontal and posterior regions of cortex were identified, suggesting 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 CS+ selective reinstatement for both 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. It is interesting that the insula showed significant ERS for extinction-specific memory, as this region was identified as selective for fear associated memory in healthy adults Overall, the searchlight analysis revealed additional cortical regions exhibiting divided organization between fear and extinction memory.

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

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 fear and extinction 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. 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 perform 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.**

Trial irrelevant “context tags” were inserted between each trial during extinction learning on day 1. These context tags can serve to track reinstatement of the encoding context during memory retrieval, as shown in previous studies (Bornstein and Norman, 2017; Gershman et al., 2013; Manning et al., 2016). Here, we explored whether evidence for the reinstatement of the extinction context during the recognition memory test predicted the bias in mPFC ERS. Specifically, does reinstated extinction context at the time of memory retrieval predict ERS in 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.

**Discussion**

Following a traumatic event, extinction learning allows an organism to form a memory of safety that acts to countervail the original fear association. An adaptive memory system will not normally overwrite threat memories with subsequent experiences of safety, such as during extinction learning. Instead, these opposing associations are stored in a way that allows for the appropriate behavior in response to ambiguous cues in the environment. Using a hybrid episodic-associative memory task we were able to describe the fine-grain organization of fear and extinction memories in the human brain. In healthy adults, we found a dissociation of fear and extinction memory storage across the mPFC and hippocampus. Specifically, extinction memories were reinstated in the vmPFC and aHC, while fear memories were reinstated in the dACC and pHC. Individuals with PTSS exhibited a similar pattern of fear memory reinstatement, but unlike healthy adults they misallocated extinction memories to regions that normally code for fear. Across both groups, we observed that various subcortical neural signals, as well as a measure of extinction context reinstatement, all predicted the bias in mPFC reinstatement between the vmPFC and dACC. These results provide strong evidence that emotional associative memories are stored in similar way in humans as they are in rodents, and provide additional support that extinction memory retrieval is dysregulated in PTSD.

The neural analyses reported here describe a general deficit in extinction memory retrieval in PTSS as compared to healthy adults. Unlike healthy adults, individuals with PTSS did not exhibit extinction related ERS in either the vmPFC or aHC. Instead, we observed a misallocation of extinction memories to regions that normally code for fear, namely the dACC and insula. These striking neural differences are unlikely to be due to any differences in prior associative learning. As discussed, we previously reported no difference in behavioral responding between healthy adults and PTSS during both fear conditioning and extinction learning. In general, there is insufficient evidence to support the idea that individuals with PTSD exhibit increased behavioral reactivity during conditioning, and subsequent impaired within-session extinction learning (Lissek and van Meurs, 2015). However, as we report in Hennings et al., 2020, individuals with PTSS did display significant fear renewal during the day 2 associative learning test, while healthy adults did not. This is in-line with impaired extinction recall being the principle behavioral deficit observed in PTSD (Lissek and van Meurs, 2015). Outside of simple discriminatory conditioning paradimg,s the dACC signals affective meaning during memory retrieval. If extinction memories are being consolidated in the dACC in PTSS, this could explain the bias to retrieve fear over safety in ambiguous situations (Kensinger and Ford, 2021). It is also worth noting that the group differences in ERS cannot be explained by differences in episodic retrieval during the recognition memory test. Both groups displayed equivalent, and overall high, recognition memory for the individual stimuli encoded during conditioning and extinction.

These results suggest that while individuals are able to form an extinction memory, they do so utilizing different neural mechanisms than healthy adults. Specifically, a failure of consolidation could result in safe associations being stored in the dACC and insula, instead of the vmPFC and aHC. We previously reported that univariate activity in the vmPFC and hippocampus was equivalent across groups during conditioning and extinction (Hennings et al., 2020a). Nevertheless, healthy adults exhibit significant reinstatement of extinction memories in these regions, while PTSS do not. If ERS were assessed immediately after extinction learning, it is possible that both healthy adults and PTSS would display similar patterns of ERS across the mPFC, with differences only emerging after a period of consolidation.

In addition to extinction memory reinstatement, in healthy adults the vmPFC also displayed significant reinstatement for items encoded during fear conditioning. Why is information about threats being represented in a region that primarily codes for learned safety? A recent study of patients with focal vmPFC lesions found an impairment of fear learning in these individuals (Battaglia et al., 2020). The authors suggest that although a main function of the vmPFC is to inhibit fear during extinction recall, it may also be necessary for the acquisition of learned fear, at least in humans. Previous research has described the role of the vmPFC in learning schemas and task rules, which may extend to discriminatory conditioning tasks. The vmPFC has been shown to signal safety outside of the framework of competing associative memories (Tashjian et al., 2021). One possibility is that reinstatement in the vmPFC reflects active safety signaling, as the stimuli encountered during the recognition no longer pose a threat. It is interesting that individuals with PTSS did not exhibit fear related ERS in the vmPFC, as it may represent a further failure of safety-related processing in this group.

The amygdala and hippocampus are critical in the learning and expression of fear and extinction. The hippocampus in particular exerts contextual control over fear and extinction, and is responsible for inhibiting the expression of extinction, leading to renewal of fear. In the present study, we found gradient of function along the long axis of the hippocampus. The posterior hippocampus (dorsal in rodents) has been shown to be necessary for both the acquisition of fear (Bast et al., 2003; Ressler et al., 2021) and extinction (Corcoran et al., 2005) in context-conditioning designs. Our results strongly suggest that the pHC is involved in the retrieval of fear memories, as both healthy adults and PTSS displayed fear conditioning specific reinstatement in the pHC. Additionally, pHC univariate activity and reinstatement both predicted a bias in mPFC reinstatement towards the dACC, which is selectively involved in the retrieval of fear.

As with the posterior region, the anterior hippocampus (ventral in rodents) has been shown to drive both the inhibition of fear (Meyer et al., 2019) and promote fear relapse (Marek et al., 2018). Our data support the idea that the aHC may be involved in both fear and extinction processes. We observed significant extinction context specific reinstatement in the aHC in healthy adults, and reinstatement predicted a bias in mPFC reinstatement to the vmPFC, which is primarily involved in extinction recall. However, we also observed that univariate activity in the anterior hippocampus predicted a bias in mPFC reinstatement to the dACC, which suggests a role in fear retrieval as well. The ability to signal both fear and extinction suggests that the aHC may be a critical region in the proper regulation of fear and safety, integrating contextual information and signaling either fear or extinction in the dACC or vmPFC, respectively. The ability of the aHC to exert bi-directional control is supported by direct projections from the aHC to both the dACC and vmFPC. Further, individuals with PTSS did not exhibit significant extinction related reinstatement in the anterior hippocampus. This observed dysregulation of the aHC may contribute to extinction recall deficits in PTSS.

In contrast to the hippocampus, we did not observe any significant CS- or context-specific reinstatement in the subfields of the amygdala. One possibility is that encountering CSs in a neutral context, such as during a recognition memory task, is sufficient to drive reinstatement in the neocortex and hippocampus, but not the amygdala. While engaged in the memory task, participants were explicitly told that there was no threat of shock, which consequently could have resulted in less threat monitoring and amygdalar processing. While rodent studies repeatedly observe amygdala activity in fear and extinction processes, direct evidence of amygdala activity in similar tasks using fMRI is scant (Fullana et al., 2016a, 2018, 2019). Even though subfields of the amygdala did not engage in significant reinstatement, we were able to show that amygdala is still involved in the balance in reinstatement between the dACC and vmPFC. Overall univariate activity in both the BLA and CeM, and local reinstatement in the CeM, all predicted a bias in mPFC reinstatement towards the dACC, consistent with the role of these regions in maintaining representations of learned fear.

Previous work utilizing multivariate pattern analyses have had some success in translating the neurobiology of fear and extinction from animal models to humnas. Pattern similarity analyses suggest that information about fear and extinction is represented in the amygdala, hippocampus, dACC, and vmPFC (Bach et al., 2011; Graner et al., 2020; Hennings et al., 2020b; Visser et al., 2013). The results presented here build on this work by providing direct evidence for neural reinstatement, i.e. engram-like activity, for fear and extinction in the human mPFC and hippocampus. An important aspect of our experimental design was that we were able to capture the reinstatement of associative memories during an orthogonal episodic memory test. The neural reinstatement of an associative memory during episodic memory retrieval is predicted by theories of episodic memory, but has not previously been shown.

This design has exciting implications for future work, as it allows experimenters to simultaneously access competing associative memories that normally exert reciprocal inhibition during associative retrieval tests. This design may be applied to future work seeking to understand the neural substrates of other associative processes, such as counter-conditioning and reconsolidation. As rodent studies have shown, the ability to attach temporally specific tags to a memory is critical for both the identification of its neural substrates, and remote control over its expression (e.g. chemogenetic stimulation). In humans, more precise identification of the neural substrates of fear and extinction may lead to better treatments of psychiatric disorders like PTSD.