**Title:**

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

**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 division exists in the human brain remains unclear. Here, we used a hybrid form of Pavlovian conditioning with an episodic memory component to identify overlapping 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; Pitman et al., 2012). 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 (Johansen et al., 2011; Liu et al., 2012; Rashid et al., 2016; Tovote et al., 2015). 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 for a particular stimulus that is stored in parallel to the memory of fear for that stimulus. 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) (Davis et al., 2017; Giustino and Maren, 2015; Herry et al., 2008; Senn et al., 2014; 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 overlapping network 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, memories change over time , and not all experiences induce a persistent change in the brain. 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. 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, 2017; 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 either express or suppress conditioned fear (Marek et al., 2018; Senn et al., 2014).

Human neuroimaging has successfully translated evidence from rodents that the dACC is among the most consistently active regions during fear conditioning (Fullana et al., 2016a). However, it is far less clear in humans whether this region is also the site of long-term storage and retrieval of the acquired fear memories, i.e. engram-like activity. 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 “label” brain activity uniquely associated with memories of either fear or 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 overlapping multivariate patterns of activity during memory encoding and retrieval. The match between patterns of activity in distributed voxels during encoding and retrieval provides an index of memory fidelity, albeit not at the cellular level. This neuroimaging technique has been widely applied to the study of human episodic 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 associative memory traces of fear and extinction in the human brain has not been tested.

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 for a memory test composed of the unique CSs encoded during both conditioning and extinction. This hybrid 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, but not both. By incorporating trial unique CSs, we can label specific episodes associated with fear and extinction by correlating patterns of activity during retrieval with patterns of activity elicited during encoding for each CS. 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 design innovation allows us to adapt technical advances in neuroimaging of human episodic memory with the conceptual framework of functional labeling from rodent neurophysiology. We hypothesized that the healthy adult brain organizes and maintains separable mnemonic representations of threat and safety, and we sought to distinguish these memories based on the temporal context in which the memory was originally formed, i.e., fear conditioning or extinction learning. Furthermore, we hypothesized that for individuals with post-traumatic stress symptoms (PTSS), fear memories would be represented similarly to healthy adults, but because of the maladaptive return of fear in PTSD, the representation of extinction memories would differ.

**RESULTS**

**Fig. 1** provides an overview of the study procedures and analytic approach. Each subject encoded trial-unique pictures of animals and tools before, during, and after 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-). Extinction learning immediately followed fear conditioning, during which no shocks were delivered. Participants returned 24-hours later for a surprise recognition memory test comprised of the CSs from each phase plus novel lures. Participants were informed that no shocks would be delivered during the memory test.



**Figure 1. Divided organization of fear and extinction memories in the human brain**. **A**. **Schematic overview.** People maintain competing representations of threat 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 them 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+s co-terminated with a mild electric shock (US). During extinction learning, a stream of natural scene images appeared between CSs (see Hennings et al., 2020). **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. To test for neural reinstatement within a given ROI, the item-specific pattern of fMRI activity elicited by each CS was correlated with the activity pattern from when that item was initially encoded. MTL: medial temporal lobe; mPFC medial prefrontal cortex; PL: prelimbic cortex; IL: infralimbic cortex; vHC: ventral hippocampus; dACC: dorsal anterior cingulate cortex; vmPFC: ventromedial prefrontal cortex; aHC: anterior hippocampus; BLA: basolateral amygdala; ITC: intercalated cells; CeA: central nucleus of the amgydala.

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**

*Explicit and implicit measures of learning*. As we reported in Hennings et al. (2020), 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). 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 (“late 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 late 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, performance on the recognition memory test replicated previous behavioral findings (Dunsmoor et al., 2015a, 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 (Hennings et al., 2021, In press). Here, we report an analysis of high confidence hit rates with the purpose of testing for a difference in episodic memory performance between healthy adults and individuals with PTSS. 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 two-way interactions between *group* and either *CS type* or *encoding context,* and no significant three-way interaction(All Ps > 0.44). These results indicate that explicit recognition memory for the CS items was equivalent between groups.

Chart

Description automatically generated

**Figure 3. Dissociable reinstatement of emotional memories. A. ERS in *a priori* ROIs.** CS+ minus CS- ERS in *a priori* dACC and vmPFC regions.Error bars correspond to the 95% confidence interval of the CS+/- difference. \*\*\*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 adults with PTSS, the dACC displayed significant emotional reinstatement of items encoded during both conditioning and extinction, revealing a misallocation of extinction memories. **B. Whole brain ERS searchlight.** A medial and lateral view of the left hemisphere is shown, and results are qualitatively similar across hemispheres. The heatmaps show average effect of CS+ minus CS- reinstatement for CS items from conditioning (red) and from extinction (blue), threshold at P < 0.001 one-sided for CS+ > CS- with a cluster-wise threshold (FWE) of P < 0.05.

**Emotional memory reinstatement in the prefrontal cortex**

The analysis here focuses on the overlap of multi-voxel fMRI activity patterns of items from encoding to retrieval (i.e., encoding retrieval similarity, ERS), irrespective of memory performance. The voxel-wise patterns of activity elicited by each CS item during the recognition memory test was correlated with the patterns of activity elicited by those same CS items when they were initially encoded during either the pre-conditioning, fear conditioning, or extinction phase. To control for item-level reinstatement effects, these correlations were Fisher z-transformed and then the average correlation of CS- trials was subtracted from the average correlation of the CS+ trials from the same encoding context. This analysis focused on two distinct subregions of the mPFC motivated by rodent work (Burgos-Robles et al., 2017; Do-Monte et al., 2015): the dACC and vmPFC. These regions were defined by drawing spheres around *a priori* peak activations (see Methods).

In healthy adults, the dACC exhibited selective reinstatement for CS+ items (compared to CS- items) that were encoded during fear conditioning (**Fig. 3**; 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. Selective reinstatement was stronger for fear memories (CS+ - CS- from conditioning) than for extinction memories (CS+ - CS- from extinction) (phase diff. = 0.21, [0.12, 0.29], P = 1.57e-6). Moreover, this region did not show any selective reinstatement of extinction memories (CS diff. = 0.014, [-0.046, 0.075], PFDR = 0.64) or pre-conditioning memories (CS diff. = 0.006, [-0.054, 0.066], P­FDR = 0.84). In sum, the dACC appears highly specialized for the reinstatement of fear memories in the healthy adult brain. In the vmPFC, there was selective reinstatement for both fear memories (CS diff. = 0.074, [0.013, 0.134], PFDR = 0.033) and extinction memories (CS diff. = 0.113, [0.053, 0.173], PFDR = 9.20e-4). There was no selective reinstatement of pre-conditioning memories (CS diff. = -0.020, [-0.081, 0.040], PFDR = 0.50). Notably, there was stronger reinstatement of fear memories in the dACC relative to the vmPFC (ROI diff. = 0.149, [0.064, 0.234], PFDR = 0.002), and stronger reinstatement of extinction memories in the vmPFC relative to the dACC (ROI diff. = 0.099, [0.014, 0.184], PFDR = 0.031). Altogether, in healthy adults, discrete regions of the mPFC exhibited a double dissociation in the selective reinstatement of fear memories and extinction memories, as identified by the temporal context in which the memories were formed.

Individuals with PTSS also exhibited selective 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 relative to the vmPFC (ROI diff. = 0.121, [0.036, 0.206], PFDR = 0.011). Unlike the healthy adult group, however, the PTSS group also showed selective reinstatement in the dACC for CS+ items encoded during extinction (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 in the PTSS group. The dACC did not exhibit selective reinstatement for pre-conditioning memories (CS diff. = 0.032, [-0.028, 0.092], PFDR = 0.30). In contrast to the healthy adult group, there was no evidence of selective reinstatement for CS+ items encoded during either conditioning (CS diff. = 0.050, [-0.010, 0.110, PFDR = 0.10) or extinction (CS diff. = 0.041, [-0.020, 0.101], PFDR = 0.19) in the vmPFC. The vmPFC did however exhibit an unexpected selectivity in the PTSS group for CS- items encoded prior to fear conditioning (-0.079, [-0.139, -0.019], PFDR = 0.024).

**Emotional memory reinstatement outside the prefrontal cortex**

To complement the results from the *a priori* ROIs, we conducted an exploratory whole brain searchlight for selective CS+ reinstatement (**Fig 3B**). In healthy adults, this analysis revealed additional brain regions exhibiting selective reinstatement of fear and extinction memories (See **Supplementary Table 1** for full list of cluster locations). In addition to the dACC, we found that fear memories were also reinstated in the anterior insula, a region consistently implicated in human fear memory (Fullana et al., 2016b). For the reinstatement of extinction memories, the largest cluster was found in vmPFC. Other cortical regions including the medial frontal gyrus and precuneus exhibited selective reinstatement for both fear and extinction memories. Individuals with PTSS were similar to healthy adults with the reinstatement of fear memories in large clusters corresponding to the dACC, bilateral insula, and other cortical regions. For extinction memories, we observed significant clusters in the cuneus, as well as bilateral insula. It is interesting that the insula showed significant reinstatement for extinction items, as this region selectively reinstated fear memories in healthy adults.

**Emotional memory reinstatement 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 memories. The hippocampus in particular exerts contextual control over memory retrieval (Maren et al., 2013). 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 (Bast et al., 2003; Corcoran et al., 2005; Marek et al., 2018; Meyer et al., 2019; Qin et al., 2021; Ye et al., 2017). Human neuroimaging also shows functional specializations for these subfields in memory and affective processes (Cooper and Ritchey, 2019; Meyer et al., 2019; Poppenk et al., 2013). Using subject-specific anatomical segmentations, we probed emotional memory reinstatement along the long-axis of the hippocampus in three bi-lateral subfields: head (anterior; aHC), body, and tail (posterior; pHC). The amygdala was similarly segmented into two bilateral ROIs known to have functional specialization in conditioning and extinction processes: the basolateral amygdala (BLA) and the central nucleus of the amygdala (CeM) (Pape and Pare, 2010).

*Hippocampus.* No selective reinstatement of CS+ items was observed for items from any encoding context in any hippocampal subfield in either healthy adults or those with PTSS (all PFDR > 0.45). However, a linear-mixed effects model revealed a significant three-way interaction of *encoding context \* subfield \* group* (X2(4­) = 12.8, P = 0.012; see **Online 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 pHC selectively reinstated items from the fear conditioning context. In healthy adults, reinstatement of fear memories was stronger than reinstatement of extinction memories (Phase diff. = 4.36e-2, [1.40e-2, 7.32e-2], PFDR = 0.019), whereas in adults with PTSS, it was stronger than both extinction memories (Phase diff. = 4.33e-2, [1.36e-2, 7.29e-2], PFDR = 0.019) and pre-conditioning memories (Phase diff. = 4.41e-2, [1.45e-2, 7.37e-2], PFDR = 0.019). The body of the hippocampus did not show reinstatement specific to any encoding context (all phase comparisons PFDR > 0.11). In contrast to the pHC, the aHC portion selectively reinstated items from extinction more than items from conditioning (ROI diff. = 0.065, [0.035, 0.094], PFDR = 3.36e-4), although this was only observed in healthy adults.

These results suggest a gradient of functional specialization along the long axis of the hippocampus. We directly tested the dissociation between the anterior and posterior regions, and found that the pHC fear exhibited more fear memory reinstatement than the aHC (ROI diff = 0.033, [0.003, 0.063], PFDR = 0.038), and the aHC exhibited more extinction memory reinstatement than the pHC (ROI diff. = 0.075, [0.046, 0.105], PFDR = 2.51e-6). (**Fig. 3**). This double dissociation held for healthy adults, but not for the PTSS group. In these participants, fear memories were biased towards the posterior hippocampus (ROI diff. = 0.034, [0.004, 0.063], PFDR = 0.038), but there was no preference along the long-axis for safety memories (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 reinstatement of emotional memories along the long axis of the hippocampus.** Reinstatement was averaged across CS+/- in each long-axis 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). \*\*\*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 subfield, and extinction reinstatement was higher in the anterior subfield. *Bottom*. In adults with 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 reinstatement of CS+ items, however none was observed for any encoding context in any subfield, in either healthy adults or those with PTSS (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.

***Recognition memory does not influence reinstatement in the mPFC.***

We additionally tested if reinstatement in our *a priori* ROIs differed as a function of memory strength. Recognition memory was included as a categorical predictor (e.g. “high-confidence hit” or “miss”)*.* In the mPFC, the was no main effect of *memory accuracy* (X2(1)­ = 0.024, P = 0.89), and all interactions with this term were not significant (all Ps > 0.14). A similar pattern emerged in the amygdala, with only a trending main effect of *memory accuracy* (X2(1) = 3.08, P = 0.08) and no significant interactions (all Ps > 0.07In the hippocampus, we again observed a trending main effect of *memory accuracy* (X2(1) = 3.47, P = 0.06), as well as several significant higher order interactions (*memory accuracy \* CS type \* encoding context*:X2(2) = 6.26, P = 0.044; *memory accuracy \* CS type \* subfield*: X2(2) = 6.69, P = 0.035; *memory accuracy \*encoding context \* group*: X2(2) = 11.8, P = 0.003).Thus, recognition memory did not influence reinstatement in the mPFC and amygdala. That recognition memory influences reinstatement in the hippocampus is not surprising given its role in episodic retrieval.

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

*Univariate activity.* Our *a priori* analysis in the mPFC showed that healthy adults exhibited a double dissociation of emotional memory 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? On a trial-by-trial basis we assessed whether neural activity levels in the same five subfields of the hippocampus and amygdala predicted the difference in ERS between our two mPFC regions (vmPFC – dACC). We restricted our analysis to CS+/- items from conditioning and extinction as our time points of interest.

We found that all subfields were significant predictors of reinstatement location, such that increases in subcortical activity was associated with reinstatement in the dACC (pHC: 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; aHC: 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 *pHC* \* *CS* *type* (X2(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 type* \* *encoding context* interaction (X2(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 reinstatement.* Next, we conducted a similar set of analyses, in which trial-by-trial reinstatement was used to predict the mPFC difference in reinstatement. 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. Item-specific memory reinstatement in three of these subfields was also predictive of reinstatement in different regions of the mPFC (**Fig. 4**). Greater reinstatement in the pHC (X2(1) = 4.64, P = 0.031, slope = -0.060) and CeM (X2(1) = 8.49, P = 0.004, slope = -0.065) was associated with a bias towards reinstatement in the dACC. In contrast, greater reinstatement in the aHC (X2(1) = 11.1, P = 8.51e-4, slope = 0.091) was associated with a bias towards reinstatement in the vmPFC. There were no significant interactions with encoding context, CS type, or group for any of these subfields. Finally, reinstatement in the body of the hippocampus and in the amygdala BLA was unrelated to reinstatement in the mPFC.

Chart

Description automatically generated

**Figure 4. Predicting the bias in mPFC reinstatement. A. Descriptive overview of the mPFC difference in memory reinstatement.** A kernel density plot shows the overall distribution of vmPFC – dACC ERS for all CSs from conditioning and extinction, collapsed across participant group. **B. Subcortical reinstatement predicts mPFC reinstatement location.** For each of the subcortical ROIs, local ERS for each trial was used to predict ERS in the mPFC. Points and error bars correspond to the estimate and 95% confidence interval of the relationship between subcortical and cortical reinstatment. \*\*\*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. Greater mean activity in the anterior hippocamps predicted a bias towards reinstatement in the dACC, while greater reinstatement in the same region predicted, a bias towards reinstatement in the vmPFC. A region which exhibits bi-directional control over cortical reinstatement of emotional memories could be crucial for the proper regulation of fear and safety in humans. Consistent with this proposition, we found that both of these factors were independently predictive of cortical reinstatement when combined into a single model (univariate: X2(1) = 42.5, P = 7.1e-11, slope = -1.65e-3; ERS: X2(1) = 5.56, P = 0.018, slope = 0.067), with no significant interactions with either predictor.

**Discussion**

Extinction learning can build a memory of safety to countervail retrieval and expression of the original fear association. However, an adaptive memory system should preserve the original fear memory, as an experience of safety does not necessarily render a stimulus completely harmless. These opposing associations should therefore be stored in a way that allows for the appropriate behavior for a given context (Moorman and Aston-Jones, 2015). Neurobiological research in rodents is beginning to reveal the structure of this organization within and between discrete brain regions by quantifying the overlap in activity during memory formation and memory expression (Letzkus et al., 2015; Reijmers et al., 2007; Tovote et al., 2015). Using multivoxel pattern similarity analysis of overlapping encoding-to-retrieval activity in human neuroimaging, we were able to identify a divided organization of fear and extinction memories in 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 selective fear memory reinstatement, but surprisingly they misallocated extinction memories to regions that were selective for associative fear memory in healthy participants. Across both groups, we observed that various subcortical neural signals that predicted the location of cortical reinstatement of emotional memories in mPFC. These results bridge an increasing collection of evidence from rodent neurophysiology for the divided organization of opposing associative memories, and provide new insights into how disorganization in these neural representations may contribute to psychiatric disease.

In healthy adults, we found a robust dissociation of fear and extinction memory reinstatement in two distinct mPFC regions. These results suggest that these regions are critical for the proper long-term storage and retrieval of opposing emotional associations. In rodents, even though the PL is not necessary for the acquisition of learned fear, it is necessary for the later expression of that fear (Corcoran and Quirk, 2007). Here, the dACC of healthy adults exhibited clear selectivity for reinstating fear memories, which confirms a role for this structure in the long-term storage of fear memories in humans. The engagement of the dACC/PL observed in the expression of learned fear likely extends beyond the storage of a simple association of CS and US. The PL receives inputs from sensory cortices, the thalamus, and other PFC regions in addition to reciprocal connections with the amygdala and hippocampus (Hoover and Vertes, 2007). These connections allow the PL to integrate information both about the external environment, and internal states to flexibly guide behavior in potentially threatening situations. In humans, the dACC is involved in the regulation emotional responses in a variety of tasks, such as signaling affective meaning during episodic memory retrieval (Kensinger and Ford, 2021). Here, the dACC was not the only cortical area to exhibit significant reinstatement of fear memories. Our whole brain searchlight analysis also revealed significant reinstatement of fear memories in the in the anterior insula, which together with the dACC are hubs of the salience network (Seeley, 2019). Thus, in ambiguous situations neural computations in and between the dACC and the other cortical described here may be critical in determining the outcome of fear retrieval.

After a fear association has been learned, the IL is necessary for the adaptive inhibition of that fear (i.e. extinction). In rodents, the IL is necessary for the long-term retention of extinction memories (Do-Monte et al., 2015), and is inhibited during fear renewal (Marek et al., 2018). Compared to the role of the dACC in fear learning and retrieval, human neuroimaging evidence for the involvement of the vmPFC in extinction is limited (Fullana et al., 2018). Here, we found that the vmPFC of healthy adults exhibited significant extinction memory reinstatement, confirming a role for this region in the long-term storage of an inhibitory extinction memory. Interestingly, we also observed that healthy adults displayed significant reinstatement of fear memories in the vmPFC. This result confirms a recent study which found an impairment of fear learning in patients with focal vmPFC lesions (Battaglia et al., 2020). That the vmPFC stores information about previous fear learning may actually be part of successful extinction learning in healthy adults. In rodents, projections from the PL to the IL are necessary for subsequent acquisition of an inhibitory association which countervails PL driven behavior (i.e. fear extinction) (Mukherjee and Caroni, 2018). Outside of the framework of competing associative memories, the vmPFC actively signals safety (Tashjian et al., 2021), and is critical for the learning of schemas and tasks rules (Hiser and Koenigs, 2018; Mack et al., 2020). Fear and extinction related reinstatement in this area may then reflect a focal example of the broader role the vmPFC/IL plays in counteracting previously learned behaviors.

In comparison to healthy adults, individuals with PTSS displayed a dysregulation of emotional memory reinstatement in the mPFC. Specifically, the dACC exhibited reinstatement of memories formed during extinction in individuals with PTSD symptoms. One possibility for this apparent misallocation is that extinction learning was deficient in these subjects. However, there was no difference in behavioral performance between healthy adults and those with PTSS during either fear conditioning or within-session extinction learning. Likewise, both groups displayed equivalent recognition memory for the stimuli encoded during conditioning and extinction. Thus, neural reinstatement appears to reflect an underlying distinction in how the extinction memory trace is formed and retrieved in these participants, and does not merely recapitulate an observable behavioral deficit. This suggests that individuals with a history of trauma may utilize an effective, but maladaptive neural mechanism for fear reduction during within-session extinction learning that bypasses the formation of a long-term extinction memory trace in the vmPFC. This idea is supported by rodent work which reported that optogenetic inhibition of the IL during extinction learning did not impair within-session extinction, but did impair later extinction recall (Do-Monte et al., 2015). Interestingly, a separate test of extinction recall (reported in Hennings et al., 2020) indeed showed greater return of extinguished fear in the PTSS group. This is consistent both with a neural mechanism of extinction learning which bypasses the vmPFC (Do-Monte et al., 2015), and previous work highlighting the extinction recall deficit often observed in PTSD (Lissek and van Meurs, 2015; Pitman et al., 2012). The misallocation of extinction-specific memories to the dACC, rather than the vmPFC, may bias the retrieval and expression of fear associations following extinction, contributing to fear relapse. It is interesting that individuals with PTSS also did not exhibit fear memory reinstatement in the vmPFC, as it may represent a further failure of safety-related processing in this group.

We also found a gradient of function along the long axis of the hippocampus, such that items were separately reinstated as a function of the temporal context of encoding, fear versus extinction. It has been shown that the hippocampus exerts contextual control over fear and extinction (O’Reilly and Rudy, 2001), and is involved with gating extinction retrieval through connections with the mPFC (Bouton et al., 2020). The posterior hippocampus (dorsal hippocampus in rodents) is necessary for both the acquisition (Bast et al., 2003; Ressler et al., 2021) and extinction (Corcoran et al., 2005) of contextual fear. Our results suggest that the pHC is involved in the retrieval of fear memories, as both healthy adults and those with PTSS displayed selective reinstatement in the pHC for items encoded during fear conditioning. 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. The anterior hippocampus (ventral hippocampus in rodents) has been shown to drive both the inhibition of fear (Meyer et al., 2019; Sierra-Mercado et al., 2011) and promote fear relapse (Marek et al., 2018; Twining et al., 2020). Our data support the idea that the aHC serves a dual role in fear and extinction. On one hand there was selective reinstatement for extinction memories in the aHC, and reinstatement in the aHC predicted reinstatement in the vmPFC. On the other hand, univariate activity in the anterior hippocampus at the time of memory retrieval predicted reinstatement to the dACC. The aHC may be critical for the proper regulation of fear and safety, integrating contextual information and gating retrieval of the fear or extinction memory through connections with 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 vmPFC (Wang et al., 2016).

Given considerable evidence of reactivation of fear engrams in the rodent basolateral amygdala (Reijmers et al., 2007), it is notable that we did not observe reinstatement in the human amygdala. One possibility is that participants were not under threat at retrieval, thereby limiting involvement of the amygdala for behavioral fear expression. However, it is worth considering that univariate amygdala activity at encoding was also weak (Hennings et al., 2020), which is consistent with meta-analyses of human fear conditioning pointing to a general lack of detectable amygdala activity with fMRI (Fullana et al., 2016a, 2018, 2019). Another possibility is that the spatial resolution limitations of fMRI are unable to separate reactivation of sparse neural population coding for both fear and extinction memories (Herry et al., 2010), as well as the CS+ and CS- (Ghosh and Chattarji, 2015). Despite the lack of observable reinstatement in the BLA, there was evidence that the amygdala helps balance selective reinstatement between the dACC and vmPFC. That is, univariate activity in both the BLA and CeM during retrieval, as well as local reinstatement in the CeM, predicted a bias in mPFC reinstatement towards the dACC. This is consistent with the idea that a distributed network of reciprocally connected brain regions organizes the storage and retrieval of fear and extinction memory (Tovote et al., 2015).

Previous work utilizing multivariate pattern analyses have had some success in translating the neurobiology of fear and extinction from animal models to humans (Bach et al., 2011; Graner et al., 2020; Hennings et al., 2020; 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. Much of the progress in the past decade on the neuroscience of fear and extinction has utilized activity-dependent functional labeling to identify the neural organization of these opposing memories and control their retrieval through chemogenetic stimulation. The results presented here extend this conceptual framework to human neuroimaging, and affords experimenters simultaneous access to observe the fidelity of competing associative memories that normally exert reciprocal inhibition during traditional tests of fear or extinction recall (e.g., spontaneous recovery). 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 may be applied to future work in humans seeking to assess the efficacy of protocols that enhance extinction (Dunsmoor et al., 2015b) or modify the underlying fear memory trace through reconsolidation updating (Ressler et al., 2021). A further possibility to extend this work is to target engagement of activity patterns unique to formation of an extinction memory in distributed networks through closed-loop decoded neurofeedback (Taschereau-Dumouchel et al., 2018, 2020) to create an enduring memory of safety. In this way, more precise localization of the neural substrates of fear and extinction may ultimately lead to better treatments of psychiatric disorders like PTSD.

**ONLINE METHODS**

**Participants***.*

A total of 48 participants from the community volunteered to complete the two-day functional MRI study. Three additional participants were recruited but did not complete the experiment. Half of the participants (N = 24; 15 female; Mean age = 21) were recruited with the criteria that they have no current or past psychiatric or neurological disorders. The remaining participants (N = 24; 17 female; Mean age = 26) were recruited after responding to flyers seeking volunteers with PTSD. These participants underwent phone screening and completed additional in-person questionnaires to confirm Criterion A trauma exposure on the PTSD checklist for DSM-5 (PCL) (Blevins et al., 2015), as well as the absence of other neurological disorders. All PTSD responding participants reported significant post-trauma symptoms related to a Criterion A trauma, however we refer to this cohort as having post-traumatic stress symptoms (PTSS) as we did not implement a structured diagnostic interview. Given high rates of co-morbid substance use disorder, all PTSS participants were given a urine toxicology screening, and no participants tested positive for illicit drugs or benzodiazepines. Written informed consent was obtained for all participants, and all experimental procedures were approved by the University of Texas at Austin IRB (#2017-02-0094). PCL scores, as well as surveys of anxiety and depression are reported in Hennings et al., 2020.

**Stimuli.**

Conditioned stimuli were images of animals and tools collected from lifeonwhite.com or other publicly available resources on the internet. Critical to the design of the task, each stimulus was a unique exemplar from its category. For example, there were not two different kinds of “dog” used. Typically phobic animals or threatening tools were excluded (e.g., spiders, snakes, knives). The unconditioned stimulus (US) was a brief (50ms) electric shock delivered to fingers of the left hand. Prior to entering the scanner, the US was calibrated for each participant to a level described as “highly annoying and unpleasant, but not painful”. A BIOPAC STMEPM-MRI module was used to deliver the US (Goleta, CA). During the recognition memory test, all 144 “old” stimuli were shown, in addition to 48 novel lures per category. CSs were presented for 3s followed by a 4 or 5s ITI (jittered). Trial order was again pseudorandomized to ensure a balance of CSs from each encoding phase as well as old and new items. Stimulus presentation was controlled using E-Prime 3.0.

**Task.**

*Associative learning task*.Participants completed an associative learning task in two sessions of about an hour each, roughly 24 hours apart. We note that “fear” is often a misnomer of the emotional construct being studied in research involving human participants (LeDoux and Pine, 2016). A better term may be “threat conditioning”*,* as it better captures both the actual emotional experience of participants and the acquisition of conditioned responses. Nevertheless, we retain the term “fear*”* to connect the results the broader field of Pavlovian conditioning. For all phases of the associative learning task, images were displayed for 4.5 +/- 0.5s (jittered), and the ITI between trials lasted 6 +/- 0.5s (jittered). The trial order of the CSs was pseudorandomized to ensure no more than 3 CS type were presented in a row. The same pseudorandomized order was used for all subjects, however which phase of the experiment each stimulus was displayed was randomized across participants. Day 1 consisted of pre-conditioning, fear conditioning, and extinction. On Day 1, each phase consisted of 48 trials, 24 animals and 24 tools, for a total of 144 items. During pre-conditioning, participants identified which category each image belonged to (2-alternative forced choice, 2-AFC; animal or tool). During fear conditioning, 50% of the trials from one category (CS+) co-terminated with the US, for a total of 12 CS+US pairings. Images from the other category were never paired with shock (CS-), and the category of the CS+ was counterbalanced across participants. Extinction learning followed fear conditioning, during which no shocks were delivered. Relevant to hypotheses explained in Hennings et al., 2020, during extinction learning the normal fixation cross displayed during the ITI was replaced with a stream of natural scene images displayed for 1s each (5, 6, or 7 scenes per ITI). During fear conditioning and extinction on Day 1, participants responded whether or not they expected a shock on each trial (2-AFC; yes or no). Skin-conductance responses were collected during pre-conditioning, fear condition, and extinction. The following day, participants had the electrodes reattached prior to entering the scanner for the fear renewal test (reported in Hennings et al., 2020).

*Recognition memory test*. After completing the fear renewal test on Day 2, participants completed a surprise recognition memory test for the items they had seen the previous day. All 144 old images were included as well as 96 novel foils. The stimuli seen during the fear renewal test were not shown during the recognition memory test. Each image was displayed for 3s with a 4 or 5s ITI, and participants indicated whether each image was old (they had seen it the previous day), or new (never seen before). Participants indicated the confidence of their choice by responding the image was definitely old, maybe old, maybe new, or definitely new. The memory test was split into three fMRI runs of equal length, and trial order was again pseudorandomized to ensure a balance of lures and foils of both CS types and encoding phases across the memory runs. Trials during the recognition memory test were removed from analysis if participants failed to make a response within the 3s window (Mean = 2.5 dropped “old” trials per participant). A perceptual localizer followed the recognition memory test to facilitate MVPA decoding, however this data was not used in the present analyses.

**Functional MRI acquisition.**

Neuroimaging was accomplished using the Siemens Skyra 3T Human MRI scanner located at the Biomedical Imaging Center at the University of Texas at Austin. Functional data were acquired with a 32-channel head-coil, with 3mm isotropic resolution (TR = 2000ms; TE = 29ms; FoV = 228; 48 slices). A multi-band factor of 2 was used with automatic AC/PC alignment. As discussed in Hennings et al., (2020), due to a computer malfunction, 2 subjects had slightly different acquisition parameters on Day 1 (TR = 2230ms; 66 slices), which were accounted for during preprocessing and analysis. An T1-weighted 3d MPRAGE scan (TR = 1900ms; 1mm isotropic resolution) was collected on Day 1 to aid in functional image registration and region of interest definition.

**Image preprocessing**

Functional MRI data were processed using *fMRIprep* (v1.5.4)*,* an open source software suite designed to increase reproducibility and develop common best practices for image processing. The following boilerplate has been included unchanged, as recommended by the package maintainers.

*Anatomical data preprocessing.* The T1-weighted (T1w) image was corrected for intensity non-uniformity (INU) 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 target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast (FSL 5.0.9, (Zhang et al., 2001). Brain surfaces were reconstructed using recon-all (FreeSurfer 6.0.1, (Dale et al., 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 one standard space (MNI152NLin2009cAsym) was performed through nonlinear registration with antsRegistration (ANTs 2.2.0), using brain-extracted versions of both T1w reference and the 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 9 BOLD runs found per subject (across all tasks and sessions), the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. Susceptibility distortion correction (SDC) was omitted as no field maps were collected. 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. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using mcflirt (FSL 5.0.9, (Jenkinson et al., 2002). BOLD runs were slice-time corrected using 3dTshift from AFNI 20160207 (Cox and Hyde, 1997). The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying the transforms to correct for head-motion. 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 run 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 and DVARS are calculated for each functional run, both using their implementations in *Nipype* (following the definitions by (Power et al., 2014). The three global signals are extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (*CompCor*, (Behzadi et al., 2007). Principal components are estimated after high-pass filtering the *preprocessed BOLD* time-series (using a discrete cosine filter with 128s cut-off) for the two *CompCor* variants: temporal (tCompCor) and anatomical (aCompCor). tCompCor components are then calculated from the top 5% variable voxels within a mask covering the subcortical regions. This subcortical mask is obtained by heavily eroding the brain mask, which ensures it does not include cortical GM regions. For aCompCor, components are calculated within the intersection of the aforementioned mask and the union of CSF and WM masks calculated in T1w space, after their projection to the native space of each functional run (using the inverse BOLD-to-T1w transformation). Components are also calculated separately within the WM and CSF masks. For each CompCor decomposition, the *k* components with the largest singular values are retained, such that the retained components’ time series are sufficient to explain 50 percent of variance across the nuisance mask (CSF, WM, combined, or temporal). The remaining components are dropped from consideration. The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. The confound time series derived from head motion estimates and global signals were expanded with the inclusion of temporal derivatives and quadratic terms for each (Satterthwaite et al., 2013). Frames that exceeded a threshold of 0.5 mm FD or 1.5 standardized DVARS were annotated as motion outliers. All resamplings can be performed with *a single interpolation step* by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and output spaces). Gridded (volumetric) resamplings were performed using antsApplyTransforms (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels (Lanczos, 1964). Non-gridded (surface) resamplings were performed using mri\_vol2surf (FreeSurfer).

**Region of interest selection**

The dACC, vmPFC, hippocampus, and amygdala were selected *a priori* to test for the presence of encoding specificity of fear and extinction memories. Prefrontal ROIs were based on peak coordinates previously reported in literature. Specifically, dACC coordinates (MNI 1, 21, 27) were taken from (Milad et al., 2007) in which a univariate contrast of CS+ > CS- during fear conditioning was used. vmPFC coordinates (MNI -4, 34, -6) were taken from an fMRI meta-analysis of extinction recall, using a univariate contrast of CS+ extinguished > CS+ unextinguished (Fullana et al., 2018). For each ROI, a sphere was drawn around the coordinates with a radius of 10mm, and was then restricted to grey matter using a grey matter probability mask with a threshold of 50%. The masks were then warped to subject space to achieve native functional resolution (3mm3) for multivariate analyses. Registration was accomplished using flirt using 12 degrees of freedom and nearest neighbor interpolation for each binary mask (FSL 5.0.9) (Jenkinso et al., 2012).

The hippocampus and amygdala were masked and segmented into subfields using Freesurfer’s segmentHA\_T1 on the preprocessed T1w anatomical images from recon-all (Freesurfer 7.0)(Fischl, 2012; Iglesias et al., 2015; Saygin et al., 2017). The hippocampus was segmented into head (anterior), body, and tail (posterior) subfields along the long axis. The amygdala was segmented into the basolateral (BLA), and central nucleus (CeM) subfields. The anatomical segmentations were registered to functional space using mri\_label2vol and binary masks created using fslmaths.

**Multivariate pattern analysis**

After preprocessing with fMRIprep, we computed a LS-S style betaseries to facilitate the encoding-retrieval similarity analysis (Mumford et al., 2012, 2014). For each scanner run, trial-specific beta images are computed iteratively using a GLM which models a single trial of interest and all other trials as regressors of no interest based on trial type (separate CS+/- no interest regressorsIn addition to the betaseries images, we also generated conventional average activity estimates for CS+ and CS- separately from each phase of learning on Day 1, (i.e. all CS+ in one regressor of interest). For GLMs of fear conditioning, the US was modeled as a 0 duration event and treated as a regressor of no interest. All GLM estimation was accomplished using FSL FEAT, prewhitening was used, and spatial smoothing was not applied in order to respect the boundaries of our *a priori* ROIs. In addition to the preprocessing applied by fMRIprep, several signals were included as confounds to be removed during GLM estimation, including the first principle component of the estimated physiological noise (aCompCor), framewise displacement, 6 standard motion parameters, and the discrete cosine-basis regressors calculated by fMRIprep for high-pass filtering.

The encoding-retrieval similarity analysis was implemented in custom Python code. The goal of this analysis was to directly compare multi-voxel patterns observed during encoding and retrieval of a specific stimulus in each ROI on a per-participant basis. In order to reduce noise prior to estimating pattern similarity, the LS-S beta images were weighted (multiplied) by the overall univariate estimate of the corresponding CS type from encoding (Hennings et al., 2020; Kim et al., 2020) (e.g. all images of extinction CS+s from encoding and retrieval were weighted by the univariate estimate of extinction CS+ activity during encoding). For each ROI, encoding-retrieval similarity was then taken as the Pearson’s correlation between the two beta images for a given stimulus, one from encoding and one from retrieval. Pearson’s r values were Fisher-z transformed and submitted to statistical analysis.

**Whole-brain searchlight.**

The searchlight analysis (Etzel et al., 2013; Kriegeskorte et al., 2006) was accomplished using the nilearn package in Python using the functional resolution images (3mm3) registered to MNI space. Images were prepared as described above, and then each pair of beta images from encoding and retrieval was submitted to a whole-brain searchlight analysis in which a Pearson’s correlation was iteratively computed in every sphere (radius = 6mm) in the brain. The resulting maps were Fisher-z transformed, and averaged by CS type and encoding context for each subject. For each context, the difference between the average CS+ - CS- maps was taken and analyzed using AFNI (v20.2.18) (Cox, 1996; Cox and Hyde, 1997; Gold et al., 1998). Specifically, 3dttest++ was used to test the CS+ - CS- difference against 0 for each encoding context and for each group. The analysis was restricted to voxels that had > 50% grey matter probability. Family-wise error correction was achieved using the -Clustsim option, which uses permutation testing to simulate the null distribution of the data in order to determine the threshold necessary to observe significant clusters. Clusters were extracted using 3dClusterize using a peak threshold of P > 0.001 (one-tailed CS+ > CS-), and a cluster threshold corresponding to P > 0.05 using full voxel connectivity. The size of the cluster necessary to reach this threshold ranged from 16-21 across the 4 maps. The coordinates of the peak voxel in each cluster were submitted to the AFNI function whereami to obtain anatomical labels based on the Talairach-Tournoux Atlas (Talairach, 1988). [something about whichever package to plot the results depending on which one I go with]

**Statistical analyses.**

With the exception of the whole-brain searchlight analysis (see above), all statistical tests are reported as two-tailed. Behavioral data was analyzed using the pingouin (Vallat, 2018) package in Python and the ez (Lawrence, 2016) package in R. As discussed in Hennings et al., (2020) due to technical errors four participants (two in each group) are missing SCR data from extinction. SCR was square-root transformed prior to analysis and analyzed using paired and independent samples t-tests (see Hennings et al., 2020 for description of SCR scoring method). 2-AFC shock expectancy from conditioning and extinction was coded as 1 = expect, and 0 = do not expect, and analyzed using paired and independent samples t-tests. As our neural analysis focused on the reinstatement of previously encoded items, the analysis of recognition memory focused on high confidence hits (i.e. definitely old responses). Hit rates were submitted to a mixed ANOVA with within subject factors of *encoding context* and *CS type*, and a between subjects factor of *group*. [something about seaborn if I let it auto-generate the 95% bootstrap CI in the plots]

All other statistical analyses were accomplished with linear mixed effects models using the *afex* (Singmann et al., 2015) package in R with maximum likelihood estimation. Encoding-retrieval similarity was analyzed on a trial wise basis, and the model included fixed effects of CS type, encoding context, subfield, and group, as well as a random intercept of subject (*ERS ~ CS type \* encoding context \* subfield \* group +* (1*|subject*)). The subfield term here represents the vmPFC/dACC when modeling ERS in the mPFC, and the subdivisions of the hippocampus for amgydala for ERS in those structures. Significance of the main effects and interactions of the fixed effects was evaluated using Chi-square tests, comparing the log-likelihoods of a model with and without the term of interest (Luke, 2017). All possible interactions were modeled, and the highest order interaction is reported for a given effect when relevant.Planned and *post-hoc* contrasts were accomplished using the emmeans (Lenth, 2019) package in R. Asymptotic degrees of freedom were used, as in general the number of observations in each model was quite large (>10,000). Parametric 95% confidence intervals are reported along with FDR corrected P-values using the p.adjust function in R. FDR correction was applied to each family of tests in each group of ROIs; for example FDR correction was applied to the 12 tests of CS+ - CS- reinstatement in the mPFC. FDR correction was also applied at the next level of analysis; for example the 4 cross-ROI comparisons of CS+ - CS- reinstatement in the mPFC.

Linear mixed-effects models were also used to evaluate whether subcortical activity predicted the difference in mPFC reinstatement (*vmPFC – dACC ERS ~ predictor \* CS type \* encoding context \* group +* (1*|subject))*. In all cases our analysis focused only on the main effect and interactions of *predictor,* which was iteratively univariate activity or local reinstatement from all subcortical ROIs. The same produce was used to evaluate the separable contributions of univariate activity and reinstatement in the aHC; both neural signals were entered as predictors in a single model. Significance of main effects and interactions was again determined using log-likelihood ratio tests, and point estimates and parametric 95% confidence intervals of the slopes were obtained using the emtrends function from emmeans.

[insert bibliography when done editing]

**Supplementary Materials**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Group | encoding context | Label (hemisphere) | mni coor.  peak | Size in voxels |
| Healthy | Acquisition | Inferior frontal gyrus (R) | 33, 9, 27 | 741 |
|  |  | Superior frontal gyrus (L) | -6. 18, 51 | 608 |
|  |  | Middle frontal gyrus (L) | -39, 33, 15 | 414 |
|  |  | Angular gyrus (L) | 33, -57, 36 | 308 |
|  |  | Insula (L) | -27, 24, -6 | 280 |
|  |  | Inferior frontal gyrus (L) | -45, 3, 21 | 178 |
|  |  | Precuneus (L) | -9, -66, 42 | 175 |
|  |  | Inferior parietal lobule (R) | 30, -54, 42 | 117 |
|  |  | Cerebellar tonsil (R) | 36, -63, -45 | 40 |
|  |  | Cerebellar tonsil (L) | -33, -60, -33 | 32 |
|  |  | Medial frontal gyrus (L) | -15, 48, -3 | 25 |
|  |  | Precuneus (R) | 12, -75, 42 | 25 |
|  |  | Middle temporal gyrus (L) | -57, -51, -6 | 22 |
|  |  |  |  |  |
|  | Extinction | Medial frontal gyrus (L) | -3, 51, 0 | 191 |
|  |  | Precuneus (L) | -6, -63, 27 | 113 |
|  |  | Angular gyrus (L) | -39, -75, 36 | 69 |
|  |  | Angular gyrus (R) | 42, -69, 30 | 44 |
|  |  | Middle temporal gyrus (R) | 63, 0, -24 | 26 |
|  |  |  |  |  |
| ptsd | Acquisition | Superior frontal gyrus (R) | -6, 18, 51 | 326 |
|  |  | Insula (L) | -33, 24, 9 | 214 |
|  |  | Insula (R) | 30, 24, -6 | 201 |
|  |  | Precuneus (L) | -18, -66, 48 | 119 |
|  |  | Supramarginal gyrus (L) | -54, -48, 27 | 62 |
|  |  | Culmen / Parahippocampal gyrus (L) | -30, -51, -24 | 61 |
|  |  | Inferior frontal gyrus (R) | 45, 6, 24 | 52 |
|  |  | Middle frontal gyrus (L) | -51, -3, 39 | 52 |
|  |  | Fusiform gyrus (L) | -57, -63, -12 | 29 |
|  |  | Precentral gyrus (L) | -42, -3, 51 | 23 |
|  |  | Middle frontal gyrus (L) | -42, 24, 24 | 22 |
|  |  | Superior temporal gyrus (L) | -51, -54, 15 | 20 |
|  |  |  |  |  |
|  | Extinction | Cuneus (L) | -6, -75, 30 | 42 |
|  |  | Insula (R) | 36, 30, 3 | 34 |
|  |  | Insula (L) | -39, 27, 0 | 25 |

**Supplementary Table 1.** Whole-brain searchlight results. Clusters correspond to significant CS+ > CS- reinstatement. Coordinates refer to the peak voxel in each cluster. Anatomical labels were derived from the Talairach-Tournoux Atlas.