Title:

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

Authors:

Augustin C. Hennings[1,2], Mason McClay[3], Michael R. Drew[1,2], Jarrod A. Lewis-Peacock[1,2,4,5], Joseph E. Dunsmoor*[1,2,5]

Affiliations:

1. Institute for Neuroscience, University of Texas at Austin, Austin, TX, USA
2. Center for Learning and Memory, Department of Neuroscience, University of Texas at Austin, Austin, TX, USA
3. Department of Psychology, University of California, Los Angeles, CA, USA
4. Department of Psychology, University of Texas at Austin, Austin, TX, USA
5. Department of Psychiatry, Dell Medical School, University of Texas at Austin, Austin, TX, USA

*Correspondence to joseph.dunsmoor@austin.utexas.edu

Summary

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 (MTL) 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., 2016). 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. Moreover, neuroimaging evidence of vmPFC involvement in extinction is 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 brain 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 two-day hybrid conditioning/episodic memory design that incorporates trial-unique (i.e., non-repeating) semantic exemplars as conditioned stimuli (CS) during fear conditioning and extinction on day 1 (Dunsmoor and Kroes, 2019). On day 2, participants undergo a memory test composed of the unique CS exemplars 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. Consequently, it is only possible to measure retrieval of either the putative fear or extinction memory at test, but not both. In our hybrid design we simultaneously isolate specific episodes associated with fear and extinction (comparable to activity-dependent labeling in murine studies) and quantify the overlap in patterns of activity for each CS as a function of the temporal context in which the CS was encoded. In this way, we can quantify whether and how these competing memories are distinctly organized in separable patterns of activity in a given participant and in a single experiment. This design innovation allows us to leverage technical advances in multivariate neuroimaging of human episodic memory within the conceptual framework of functional labeling from rodent neurophysiology.

We hypothesized that the healthy adult brain organizes and maintains separable mnemonic representations of fear and extinction, and we sought to distinguish these memories based on the temporal context in which the memory was originally formed. We hypothesized that

fear memories would be represented similarly in healthy adults and individuals with post-traumatic stress symptoms (PTSS). However, based on extensive evidence of maladaptive return of fear in PTSD, we hypothesized that neural organization of extinction memories would differ between groups.

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 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.

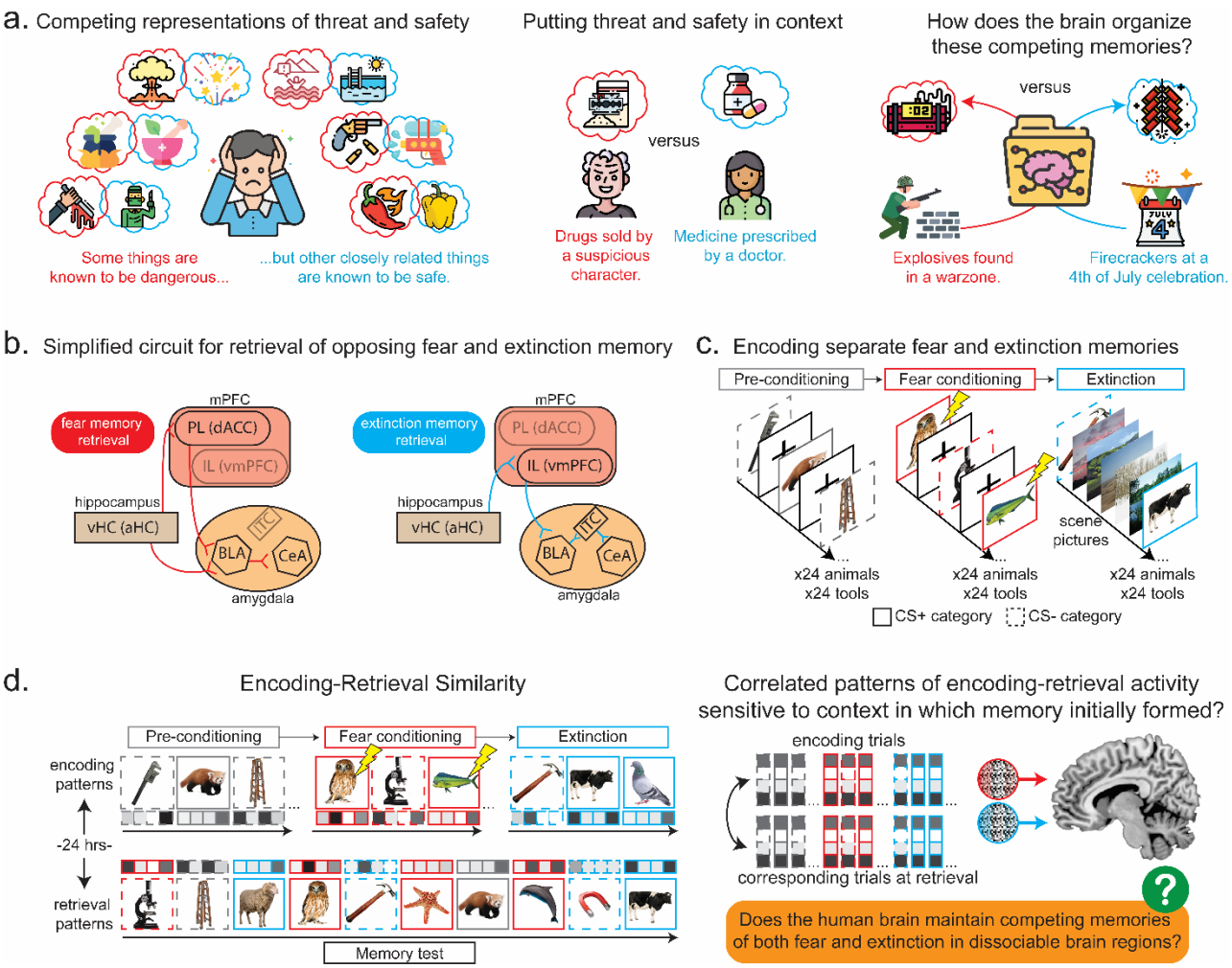


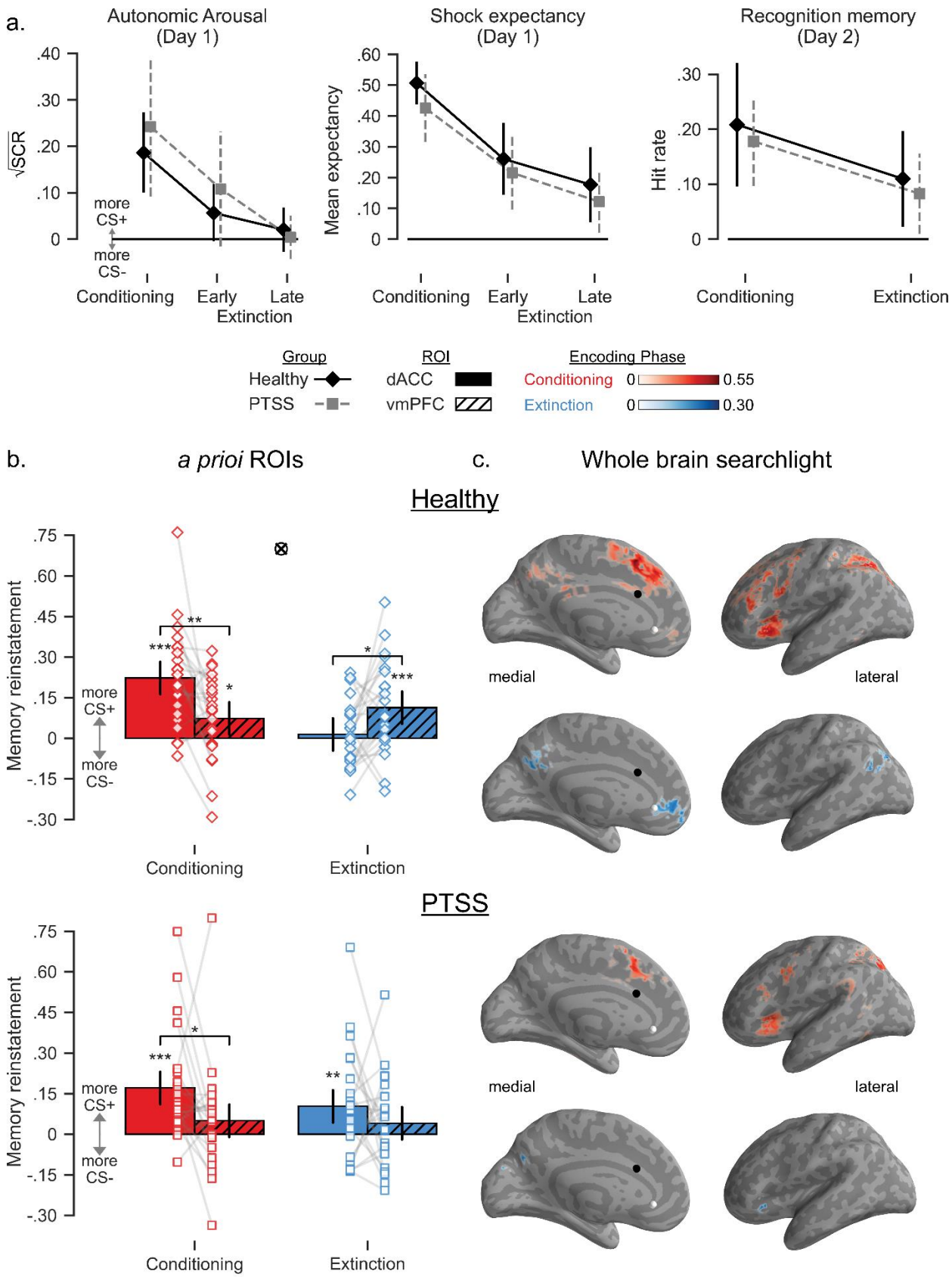
Figure 1. Divided organization of fear and extinction memories in the human brain. **a. Schematic overview.** People maintain competing representations of threat and safety for closely related stimuli or situations. We can often retrieve the appropriate associative memory given the context. How the brain 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. 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 amygdala.

Behavioral results

Explicit and implicit measures of learning. As we previously reported (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; AFC). Analyses focused on differential responding (i.e., CS+ > CS- differences) in SCR and shock expectancy from each phase (**Fig. 2a**). 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) between these two phases. 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. 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* ($F_{1, 46}$ = 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 not different between groups.


a.
Autonomic Arousal (Day 1)
√SCR
more CS+
more CS-
Conditioning
Early Extinction
Late Extinction
Shock expectancy (Day 1)
Mean expectancy
Recognition memory (Day 2)
Hit rate
Conditioning
Extinction
Group
Healthy
PTSS
ROI
dACC
vmPFC
Encoding Phase
Conditioning 0 0.55
Extinction 0 0.30
b.
a prioi ROIs
c.
Whole brain searchlight
Healthy
Memory reinstatement
medial
lateral
PTSS

170

Figure 2. Dissociable reinstatement of emotional memories. All error bars correspond to the 95% confidence interval of the CS+ – CS- difference. ***P < 0.001, **P < 0.01, *P < 0.05, FDR corrected. **a. Behavioral responses.** Autonomic arousal and explicit shock expectancy from fear conditioning and extinction (split by 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. 24hr delayed recognition memory hit rates are shown for items encoded during learning. Critically, no group differences were observed in behavioral responses during associative learning or the recognition memory test. See text for statistical analyses. **b. Reinstatement in *a priori* ROIs.** *Top.* Healthy adults exhibited a significant 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 analysis.** A searchlight was run over the entire brain, calculating reinstatement in each location. Medial and lateral views of the inflated left hemisphere are shown; results were qualitatively similar across hemispheres. The heatmaps show average CS+ – CS- reinstatement for items from conditioning (red) and from extinction (blue). Maps were threshold at P < 0.001 one-sided for CS+ > CS- with a cluster-wise threshold (FWE) of P < 0.05. The centers of *a priori* ROIs are marked on the cortical surface for the dACC (black) and vmPFC (white).

Emotional memory reinstatement in the medial 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), 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 **Online Methods**).

In healthy adults, the dACC exhibited selective reinstatement for CS+ items (compared to CS- items) that were encoded during fear conditioning (**Fig. 2b, top**; difference = 0.22, 95% CI = [0.16, 0.28], P_{FDR} = 4.62e-12). This finding accords with rodent models that show the PL is

involved in both the learning and retrieval of long-term fear memories. Selective reinstatement in the dACC was stronger for fear memories (CS+ – CS- from conditioning) than for extinction memories (CS+ – CS- from extinction; 0.21, [0.12, 0.29], P_{FDR} = 6.26e-6). Moreover, this region did not show any selective reinstatement of extinction memories (0.014, [-0.046, 0.075], P_{FDR} = 0.64) or pre-conditioning memories (0.006, [-0.054, 0.066], P_{FDR} = 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 (0.074, [0.013, 0.134], P_{FDR} = 0.033) and extinction memories (0.113, [0.053, 0.173], P_{FDR} = 9.20e-4). There was no selective reinstatement of pre-conditioning memories (-0.020, [-0.081, 0.040], P_{FDR} = 0.50). Notably, there was a significant double dissociation in the selective reinstatement of fear and extinction memories between these two regions (significant *CS type * encoding context * ROI* interaction; $X^2_{(1)}$ = 16.2, P = 5.71e-5). Specifically, there was stronger reinstatement of fear memories in the dACC relative to the vmPFC (0.149, [0.064, 0.234], P_{FDR} = 0.002), and stronger reinstatement of extinction memories in the vmPFC relative to the dACC (0.099, [0.014, 0.184], P_{FDR} = 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.

As with healthy adults, individuals with PTSS also exhibited selective reinstatement of CS+ items in the dACC for items encoded during conditioning (**Fig. 2b**, bottom; 0.171, [0.111, 0.231], P_{FDR} = 1.53e-7), and reinstatement of fear memories was stronger in the dACC relative to the vmPFC (0.121, [0.036, 0.206], P_{FDR} = 0.011). The dACC also did not exhibit selective reinstatement for pre-conditioning memories (0.032, [-0.028, 0.092], P_{FDR} = 0.30). This pattern of selective fear memory reinstatement is consistent with results in healthy adults and suggests that individuals with PTSS do not exhibit a fear processing related deficit. Unlike the healthy adult group, however, the PTSS group also showed selective reinstatement for CS+ items encoded

during extinction in the dACC (0.103, [0.043, 0.164], P_{FDR} = 0.002). These results suggest that individuals with PTSS are misallocating extinction memories, as information encoded in the extinction context was reinstated in the same region involved in the formation and retrieval of fear memories. In the vmPFC, there was unexpected selectivity for CS- items encoded prior to fear conditioning (-0.079, [-0.139, -0.019], P_{FDR} = 0.024). In contrast to the healthy adult group, there was no evidence of selective reinstatement for CS+ items encoded during either conditioning (0.050, [-0.010, 0.110, P_{FDR} = 0.10) or extinction (0.041, [-0.020, 0.101], P_{FDR} = 0.19) in the vmPFC. The significant double dissociation of fear and extinction memory reinstatement we observed in the healthy adults was not present in the PTSS group (no significant *CS type * encoding context * ROI* interaction; $X^2_{(1)}$ = 0.88, P = 0.35). Thus, while individuals with PTSS exhibit normal reinstatement of fear memories in the dACC, this group did not exhibit any selective reinstatement of CS+ memories in the vmPFC. Instead, extinction memory reinstatement was misallocated to the dACC.

Emotional memory reinstatement outside *a priori* cortical ROIs

To complement the results from the *a priori* ROIs, we conducted an exploratory whole brain searchlight for selective CS+ reinstatement (**Fig. 2c**). 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 reinstated in the anterior insula, a region consistently implicated in human fear memory (Fullana et al., 2016). 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 in 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 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 $P_{FDR} \geq 0.45$). However, a linear-mixed effects model revealed a significant three-way interaction of *encoding context * subfield * group* ($X^2_{(4)} = 12.8$, $P = 0.012$; 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 reinstatement by encoding context, collapsing 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 (4.36e-2, [1.40e-2, 7.32e-2], $P_{FDR} = 0.019$), whereas in

adults with PTSS, it was stronger than both extinction memories (4.33e-2, [1.36e-2, 7.29e-2], P_{FDR} = 0.019) and pre-conditioning memories (4.41e-2, [1.45e-2, 7.37e-2], P_{FDR} = 0.019). The body of the hippocampus did not show reinstatement specific to any encoding context (all phase comparisons $P_{FDR} \geq 0.11$). In contrast to the pHC, the aHC portion selectively reinstated items from extinction more than items from conditioning (0.065, [0.035, 0.094], P_{FDR} = 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 aHC and pHC subfields and found a significant double dissociation in healthy adults (significant *encoding context * subfield* interaction; $X^2_{(1)}$ = 23.04, P = 1.59e-6). Specifically, the pHC exhibited more fear memory reinstatement than the aHC (0.033, [0.003, 0.063], P_{FDR} = 0.038), and the aHC exhibited more extinction memory reinstatement than the pHC (0.075, [0.046, 0.105], P_{FDR} = 2.51e-6). (**Fig. 3**). This double dissociation was not observed in the PTSS group (no significant *encoding context * subfield* interaction; $X^2_{(1)}$ = 1.80, P = 0.19). In these participants, fear memories were biased towards the pHC (0.034, [0.004, 0.063], P_{FDR} = 0.038), but there was no preference between the aHC and pHC for extinction memories (-0.004, [-0.034, 0.026], P_{FDR} = 0.80). The lack of extinction reinstatement in the aHC further supports the idea that the neural organization of safety memories is dysregulated in PTSS as compared to healthy adults.

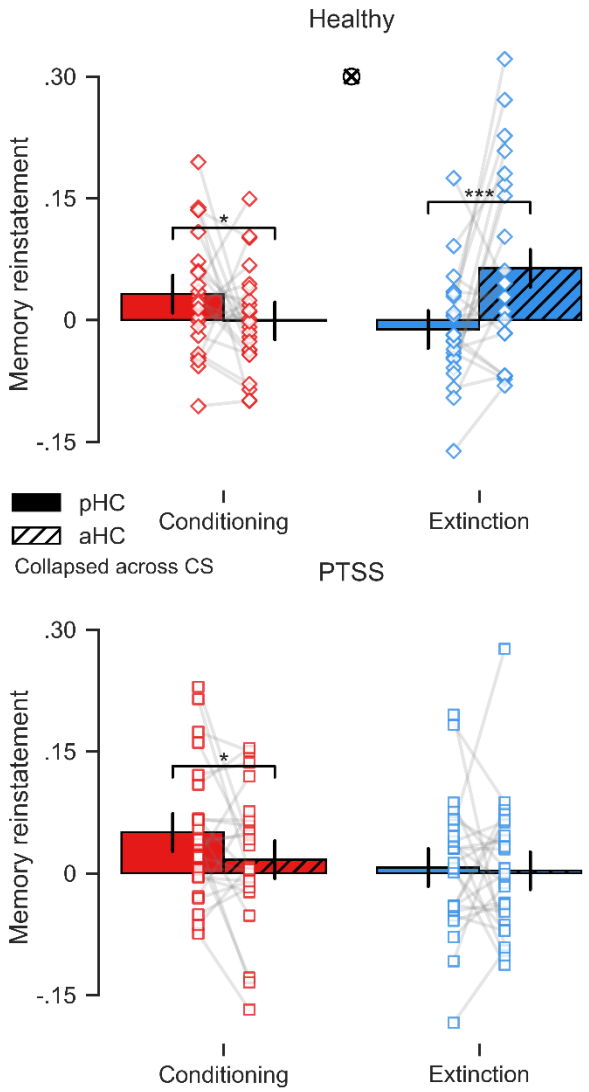


Figure 3. Differential reinstatement of emotional memories along the long axis of the hippocampus. Reinstatement was collapsed across CS+/- by encoding context in each long-axis subfield of the hippocampus. Error bars correspond to the 95% confidence interval of the marginal means. Phase specific reinstatement was observed in the pHC and aHC 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 pHC, and extinction reinstatement was higher in the aHC. *Bottom*. In adults with PTSS, the pHC subfield exhibited more reinstatement of conditioning items than the aHC.

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 $P_{FDR} \geq 0.64$). 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.

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. What determines in which area of the mPFC a particular item is reinstated? The hippocampus and amygdala both contain discrete but spatially intermixed populations of neurons that code for fear and extinction (Lacagnina et al., 2019; Senn et al., 2014; Zhang et al., 2020). Shifts in activity between these populations balances the behavioral expression of emotional memories, in part through their differing long-range connections with the mPFC (Marek et al., 2018; Senn et al., 2014; Sotres-Bayon et al., 2012). Regions that exhibit bi-directional control over the expression of emotional memories could be crucial for the proper regulation of fear and extinction in humans. Thus, on a trial-by-trial basis, we assessed whether neural activity levels in the subfields of the hippocampus and amygdala predicted the location of reinstatement between our two mPFC regions (vmPFC and 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 MTL activity at the time of memory retrieval predicted more reinstatement in the dACC (pHC: $X^2_{(1)} = 54.7$, P = 1.38e-13, slope = -1.8e-3; HC body: $X^2_{(1)} = 68.2$, P = 1.48e-16, slope = -2.49e-3; aHC: $X^2_{(1)} = 46.8$, P = 8.00e-12, slope = -1.7e-3; BLA: $X^2_{(1)} = 26.7$, P = 2.39e-7, slope = -1.45e-3; CeM: $X^2_{(1)} = 19.5$, P = 1.01e-5, slope = -6.90e-4). Additionally, we observed several interactions with the hippocampal subfields, see **Supplementary Results**.

MTL reinstatement. Having established that overall activity in the MTL predicts more reinstatement in the dACC, we next conducted a similar set of analyses in which trial-by-trial reinstatement was used to predict the mPFC difference in reinstatement. The hypothesis for what

MTL reinstatement will predict is not automatically the same as what was observed for univariate activity, as the information present in a spatial pattern of activity differs from the mean activity across that pattern. Item-specific memory reinstatement in three subfields was predictive of reinstatement in different regions of the mPFC (**Fig. 4**). Greater reinstatement in the pHC ($X^2_{(1)}$ = 4.64, P = 0.031, slope = -0.060) and CeM ($X^2_{(1)}$ = 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 ($X^2_{(1)}$ = 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 the BLA did not predict mPFC reinstatement location.

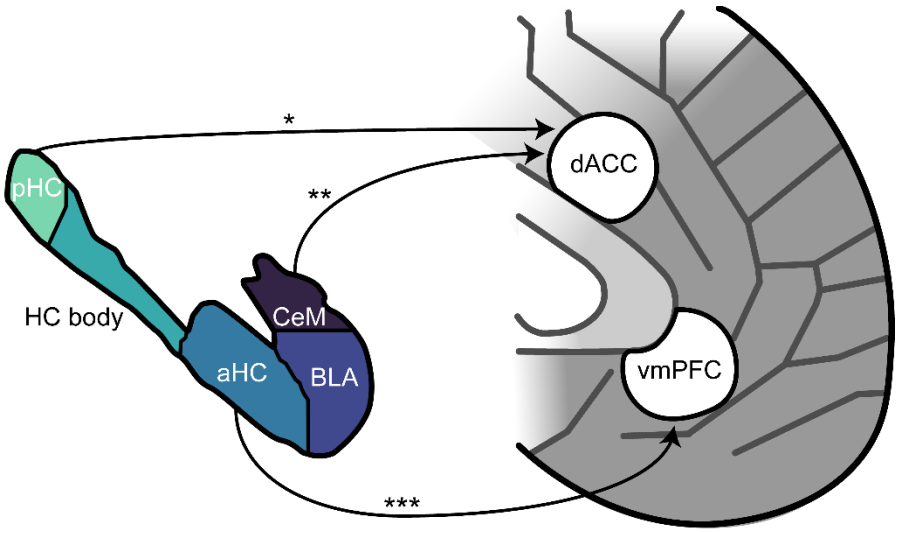


Figure 4. MTL reinstatement predicts mPFC reinstatement location. For each of the ROIs in the hippocampus and amygdala, local reinstatement was used to predict the difference in reinstatement in the mPFC. A stylized representation of the MTL and mPFC is shown. Arrows indicate a significant prediction, the direction of which was determined by the slope; see text for details. ***P < 0.001, **P < 0.01, *P < 0.05.

Separable influence of the aHC. We found that univariate and multivariate signals from the aHC predict opposite biases in mPFC reinstatement during memory retrieval. Greater mean activity in the aHC predicted a bias in reinstatement to the dACC, while greater reinstatement in

the same region predicted a bias to the vmPFC. Consistent with the proposition that the aHC exerts bidirectional control over the expression of fear and extinction, we found that both neural signatures were independently predictive of cortical reinstatement when combined into a single model (univariate: $X^2_{(1)}$ = 42.5, P = 7.1e-11, slope = -1.65e-3; reinstatement: $X^2_{(1)}$ = 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. However, they surprisingly misallocated extinction memories to regions selective for associative fear memory in healthy participants. Across both groups, we observed that various neural signals from the MTL predicted the location of cortical reinstatement of emotional memories in mPFC. These results bridge increasing 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.

Previous findings from rodents show the PL is necessary for the long-term retrieval and expression of conditioned fear (Corcoran and Quirk, 2007; Giustino and Maren, 2015). 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 from the external environment as well as internal states to flexibly guide behavior in potentially threatening situations. Here, we found that the dACC reinstates activity patterns unique to the formation of associative fear memories, confirming a role for this structure in the organization of long-term fear memories in the human brain. The dACC is broadly activated by negative emotional stimuli (Etkin et al., 2011), including during emotional episodic memory retrieval (Kensinger and Ford, 2021). A whole brain searchlight analysis also revealed reinstatement of fear memories in the anterior insula, which together with the dACC are hubs of the salience network (Seeley, 2019). Collectively, fear memory representations appear distributed across cortical and subcortical networks that may code for unique aspects of the fear experience (Frankland et al., 2019; Wheeler et al., 2013). Orchestration between these regions likely determines retrieval of fear memory over extinction memory.

The rodent IL is necessary for the long-term retention of extinction memories (Do-Monte et al., 2015), and is inhibited by the ventral hippocampus during fear renewal (Marek et al., 2018). Here, we found that the vmPFC reinstates activity patterns unique to the formation of an extinction memory in the healthy adult brain. Notably, univariate human neuroimaging evidence for the involvement of the vmPFC in extinction learning and recall has been limited and mixed (Fullana et al., 2018). The present finding thus helps bridge extensive evidence from rodents to humans on the role of this region in organizing extinction memory to inhibit retrieval and expression of fear associations. We also found that individuals with PTSD symptoms displayed an abnormal organization of fear and extinction reinstatement in the mPFC. Specifically, the dACC exhibited reinstatement of memories formed during extinction, while there was an absence of selective

extinction memory reinstatement in the vmPFC. Critically, behavioral performance of within-session extinction learning was not different between groups, and both groups remembered an equivalent number of items encoded during extinction. Thus, distinctions in neural reinstatement of extinction memory appear to reflect an underlying distinction in how an implicit extinction memory trace is formed and retrieved; they do not merely recapitulate an observable behavioral deficit. This suggests that individuals with a history of trauma may utilize a different, and ultimately 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. Interestingly, evidence from rodent studies shows the IL is not necessarily required for within-session extinction, only for successful extinction retrieval (Do-Monte et al., 2015). However stimulation of the vmPFC during or after extinction learning improves extinction retention (Do-Monte et al., 2015; Dunsmoor et al., 2019; Haaker et al., 2013; Milad et al., 2004; Raij et al., 2018). The inability to form an initial memory trace of extinction learning in the vmPFC, to be retrieved at a later time, may therefore be a critical factor in extinction retrieval deficits observed in PTSD (Lissek and van Meurs, 2015; Pitman et al., 2012). Likewise, the misallocation of extinction-specific memories to the dACC, rather than the vmPFC, may bias the retrieval and expression of fear associations following extinction, further contributing to fear relapse. These provide potential targets to strengthen extinction memory for clinical purposes.

We also found divided organization of fear and extinction memories along the long axis of the hippocampus. Interestingly, neural reinstatement in the hippocampus was sensitive to the temporal context of encoding (fear versus extinction) rather than the valence of the CS (CS+ versus CS-). This division of contextual specificity aligns with the general role of the hippocampus in forming contextual representation in associative learning (O'Reilly and Rudy, 2001) and exerting contextual control of extinction retrieval through connections with the mPFC (Bouton et al., 2020; Marek et al., 2018). The hippocampus maintains competing representations of fear and

extinction memory in distinct neural populations in the dentate gyrus (Lacagnina et al., 2019) and CA1 (Tronson et al., 2009). Whether there is a division in dorsal and ventral regions in the representation of fear versus extinction memory is less clear. This organization is likely determined by dissociable connectivity with subregions of the mPFC (Bast et al., 2003; Corcoran et al., 2005; Marek et al., 2018; Meyer et al., 2019; Qin et al., 2021; Sierra-Mercado et al., 2011; Twining et al., 2020; Ye et al., 2017). 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, neural reinstatement in the pHC, as well as univariate activity at the time of retrieval, predicted a bias in mPFC reinstatement towards the dACC. The aHC, in contrast, showed selective reinstatement for items encoded in the extinction context. However, further analysis showed that the aHC serves a dual role in retrieval of fear and extinction memory. On one hand, neural reinstatement in the aHC predicted neural reinstatement in the vmPFC, suggesting a network for extinction memory organization. On the other hand, univariate activity in the aHC at the time of memory retrieval predicted reinstatement in the dACC, consistent with a separate network that may facilitate retrieval of associative fear memories. The aHC therefore appears well situated for integrating contextual information and gating retrieval of the fear or extinction memory through connections with the dACC or vmPFC, respectively.

Given considerable evidence of reactivation of fear engrams in the rodent basolateral amygdala (e.g., (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, there was a general lack of amygdala involvement at encoding as well, consistent with meta-analyses of fMRI of human fear conditioning (Fullana et al., 2016, 2018, 2019). The spatial resolution limitations of fMRI are perhaps 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). Although we did not observe selective reinstatement in the amygdala, univariate activity in the amygdala during retrieval, as well as local reinstatement in the CeM, predicted reinstatement in the dACC over the vmPFC (**Fig. 4**). This is consistent with the idea that reciprocal connections between the amygdala and mPFC organizes the storage and retrieval of fear memories (Tovote et al., 2015).

To conclude, 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 (Frankland et al., 2019; Tovote et al., 2015). Here we provide evidence for selective neural reinstatement of fear and extinction memory representations in the human brain through overlapping activity patterns at encoding and retrieval. These results extend a conceptual framework of engram-like representations, and more broadly bolsters the use of multivariate pattern analyses to translate cutting-edge advances in the neurobiology of fear and extinction to humans (Bach et al., 2011; Graner et al., 2020; Hennings et al., 2020; Visser et al., 2013). The hybrid episodic/conditioning design incorporated here afforded us simultaneous access to isolate memories that normally exert reciprocal inhibition during a traditional test of an extinguished memory (e.g., spontaneous recovery or renewal test). Selective neural reinstatement of competing memories formed under different temporal contexts is predicted by the encoding-specificity principle (Tulving and Thomson, 1973) and neural reinstatement of episodic memory in human neuroimaging (Polyn et al., 2005; Ritchey et al., 2013), but has not previously been shown for fear and extinction memory in humans. 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 networks involved in organizing fear and extinction memory could 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” can be 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 across model organisms. 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. Participants were informed that no shocks would be delivered during the memory test. 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 preprocessnoning. 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 (Fullana et al., 2018), using a univariate contrast of CS+ extinguished > CS+ unextinguished. 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 ($3mm^3$) 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 (Jenkinson 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 general linear model (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 regressors). In 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 ($3mm^3$) 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 encoding 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). The *pysurfer* package in Python was used to resample and slightly smooth (FWHM = 1mm) the cluster maps onto the cortical surface for display purposes.

Statistical analyses.

With the exception of the whole-brain searchlight analysis (see above), all statistical tests are reported as two-tailed, and all estimates of error are given as parametric 95% confidence intervals (i.e., 1.96 * standard error of the mean). 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*

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 (*reinstatement ~ CS type * encoding context * subfield * group* + (1|*subject*)). The subfield term here represents the vmPFC/dACC when modeling reinstatement in the mPFC, and the subdivisions of the hippocampus and amygdala for reinstatement 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. When testing the double dissociations of reinstatement in the mPFC and hippocampus, data was restricted to CS items from conditioning and extinction, and a separate model was fit for each group (without the *group* term). All 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 (between ~4,000 up to ~20,000). Parametric 95% confidence intervals of the differences 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 (2 ROIs, 3 contexts, 2 groups). 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 (2 ROIs, 2 groups).

Linear mixed-effects models were also used to evaluate whether MTL activity predicted the difference in mPFC reinstatement (*vmPFC – dACC reinstatement ~ 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 MTL ROIs. The same procedure 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*.

AUTHOR CONTRIBUTIONS

A.C.H., J.E.D, and J. A. L.P., conceived of and designed the fMRI experiment. A.C.H. and M. M. implemented the fMRI experiment and collected the data. A.C.H preprocessed and analyzed the data and visualized results. A.C.H, J.E.D., and J.A.L.P. wrote the original draft of the manuscript. A.C.H, M.R.D., J.E.D., and J.A.L.P. reviewed and edited the final draft of the manuscript.

COMPETING INTEREST STATEMENT

The authors declare that they have no competing interests.

DATA AVAILABILITY STATEMENT

All deidentified neuroimaging and behavioral data may be found online at https://osf.io/qeg83/

CODE AVAILABILITY STATEMENT

All custom python and R code used for analysis will be made available at https://osf.io/qeg83/

REFERENCES

Avants, B.B., Epstein, C.L., Grossman, M., and Gee, J.C. (2008). Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. Medical Image Analysis *12*, 26–41.

Bach, D.R., Weiskopf, N., and Dolan, R.J. (2011). A stable sparse fear memory trace in human amygdala. Journal of Neuroscience *31*, 9383–9389.

Bast, T., Zhang, W.N., and Feldon, J. (2003). Dorsal hippocampus and classical fear conditioning to tone and context in rats: Effects of local NMDA-receptor blockade and stimulation. Hippocampus *13*, 657–675.

Behzadi, Y., Restom, K., Liau, J., and Liu, T.T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. NeuroImage *37*, 90–101.

Blevins, C.A., Weathers, F.W., Davis, M.T., Witte, T.K., and Domino, J.L. (2015). The Posttraumatic Stress Disorder Checklist for DSM-5 (PCL-5): Development and Initial Psychometric Evaluation. Journal of Traumatic Stress *28*, 489–498.

Bouton, M.E., Maren, S., and McNally, G.P. (2020). Behavioral and neurobiological mechanisms of pavlovian and instrumental extinction learning. Physiological Reviews *101*, 611–681.

Burgos-Robles, A., Vidal-Gonzalez, I., and Quirk, G.J. (2009). Sustained conditioned responses in prelimbic prefrontal neurons are correlated with fear expression and extinction failure. Journal of Neuroscience *29*, 8474–8482.

Burgos-Robles, A., Kimchi, E.Y., Izadmehr, E.M., Porzenheim, M.J., Ramos-Guasp, W.A., Nieh, E.H., Felix-Ortiz, A.C., Namburi, P., Leppla, C.A., Presbrey, K.N., et al. (2017). Amygdala inputs to prefrontal cortex guide behavior amid conflicting cues of reward and punishment. Nature Neuroscience *20*, 824–835.

Cooper, R.A., and Ritchey, M. (2019). Cortico-hippocampal network connections support the multidimensional quality of episodic memory. ELife *8*.

Corcoran, K.A., and Quirk, G.J. (2007). Activity in Prelimbic Cortex Is Necessary for the Expression of Learned, But Not Innate, Fears. Journal of Neuroscience *27*, 840–844.

Corcoran, K.A., Desmond, T.J., Frey, K.A., and Maren, S. (2005). Hippocampal inactivation disrupts the acquisition and contextual encoding of fear extinction. Journal of Neuroscience *25*, 8978–8987.

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

Cox, R.W., and Hyde, J.S. (1997). Software tools for analysis and visualization of fMRI data. NMR Biomed *10*, 171–178.

Dale, A.M., Fischl, B., and Sereno, M.I. (1999). Cortical surface-based analysis: I. Segmentation and surface reconstruction. NeuroImage *9*, 179–194.

Davis, P., Zaki, Y., Maguire, J., and Reijmers, L.G. (2017). Cellular and oscillatory substrates of fear extinction learning. Nat Neurosci *20*, 1624–1633.

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

Dunsmoor, J.E., and Kroes, M.C. (2019). Episodic memory and Pavlovian conditioning: ships passing in the night. Current Opinion in Behavioral Sciences *26*, 32–39.

Dunsmoor, J.E., Murty, V.P., Davachi, L., and Phelps, E.A. (2015a). Emotional learning selectively and retroactively strengthens memories for related events. Nature *520*, 345–348.

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

Dunsmoor, J.E., Kroes, M.C.W., Moscatelli, C.M., Evans, M.D., Davachi, L., and Phelps, E.A. (2018). Event segmentation protects emotional memories from competing experiences encoded close in time. Nature Human Behaviour *2*, 291–299.

Dunsmoor, J.E., Kroes, M.C.W., Li, J., Daw, N.D., Simpson, H.B., and Phelps, E.A. (2019). Role of human ventromedial prefrontal cortex in learning and recall of enhanced extinction. The Journal of Neuroscience 2713–2718.

Etkin, A., Egner, T., and Kalisch, R. (2011). Emotional processing in anterior cingulate and medial prefrontal cortex. Trends in Cognitive Sciences *15*, 85–93.

Etzel, J.A., Zacks, J.M., and Braver, T.S. (2013). Searchlight analysis: Promise, pitfalls, and potential. NeuroImage *78*, 261–269.

Fischl, B. (2012). FreeSurfer. NeuroImage *62*, 774–781.

Fonov, V., Evans, A., McKinstry, R., Almli, C., and Collins, D. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. NeuroImage *47*, S102.

Frankland, P.W., Josselyn, S.A., and Köhler, S. (2019). The neurobiological foundation of memory retrieval. Nature Neuroscience *22*, 1576–1585.

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

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

Fullana, M.A., Albajes-Eizagirre, A., Soriano-Mas, C., Vervliet, B., Cardoner, N., Benet, O., Radua, J., and Harrison, B.J. (2019). Amygdala where art thou? Neuroscience and Biobehavioral Reviews *102*, 430–431.

Ghosh, S., and Chattarji, S. (2015). Neuronal encoding of the switch from specific to generalized fear. Nature Neuroscience *18*, 112–120.

Giustino, T.F., and Maren, S. (2015). The Role of the Medial Prefrontal Cortex in the Conditioning and Extinction of Fear. Front. Behav. Neurosci. *9*.

Gold, S., Christian, B., Arndt, S., Zeien, G., Cizadlo, T., Johnson, D.L., Flaum, M., and Andreasen, N.C. (1998). Functional MRI statistical software packages: a comparative analysis. Hum Brain Mapp *6*, 73–84.

Graner, J.L., Stjepanović, D., and LaBar, K.S. (2020). Extinction learning alters the neural representation of conditioned fear. Cognitive, Affective and Behavioral Neuroscience.

Greve, D.N., and Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. NeuroImage *48*, 63–72.

Haaker, J., Gaburro, S., Sah, A., Gartmann, N., Lonsdorf, T.B., Meier, K., Singewald, N., Pape, H.-C., Morellini, F., and Kalisch, R. (2013). Single dose of L-dopa makes extinction memories context-independent and prevents the return of fear. Proceedings of the National Academy of Sciences *110*, E2428–E2436.

Hennings, A.C., McClay, M., Lewis-Peacock, J.A., and Dunsmoor, J.E. (2020). Contextual reinstatement promotes extinction generalization in healthy adults but not PTSD. Neuropsychologia *147*, 107573.

Herry, C., Ciocchi, S., Senn, V., Demmou, L., Müller, C., and Lüthi, A. (2008). Switching on and off fear by distinct neuronal circuits. Nature *454*, 600–606.

Herry, C., Ferraguti, F., Singewald, N., Letzkus, J.J., Ehrlich, I., and Lüthi, A. (2010). Neuronal circuits of fear extinction. European Journal of Neuroscience *31*, 599–612.

Hoover, W.B., and Vertes, R.P. (2007). Anatomical analysis of afferent projections to the medial prefrontal cortex in the rat. Brain Struct Funct *212*, 149–179.

Iglesias, J.E., Augustinack, J.C., Nguyen, K., Player, C.M., Player, A., Wright, M., Roy, N., Frosch, M.P., McKee, A.C., Wald, L.L., et al. (2015). A computational atlas of the hippocampal formation using ex vivo , ultra-high resolution MRI: Application to adaptive segmentation of in vivo MRI. NeuroImage *115*, 117–137.

Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved Optimization for the Robust and Accurate Linear Registration and Motion Correction of Brain Images. NeuroImage *17*, 825–841.

Jenkinson, M., Beckmann, C., Behrens, T., Woolrich, M., and Smith, S. (2012). FSL. Neuroimage *62*, 782–790.

Johansen, J.P., Cain, C.K., Ostroff, L.E., and LeDoux, J.E. (2011). Molecular Mechanisms of Fear Learning and Memory. Cell *147*, 509–524.

Johnson, J.D., McDuff, S.G.R., Rugg, M.D., and Norman, K.A. (2009). Recollection, Familiarity, and Cortical Reinstatement: A Multivoxel Pattern Analysis. Neuron *63*, 697–708.

Josselyn, S.A., Köhler, S., and Frankland, P.W. (2015). Finding the engram. Nat Rev Neurosci *16*, 521–534.

Keller, N.E., and Dunsmoor, J.E. (2020). The effects of aversive-to-appetitive counterconditioning on implicit and explicit fear memory. Learning & Memory *27*, 12–19.

Kensinger, E.A., and Ford, J.H. (2021). Guiding the Emotion in Emotional Memories: The Role of the Dorsomedial Prefrontal Cortex. Curr Dir Psychol Sci 0963721421990081.

Kim, H., Smolker, H.R., Smith, L.L., Banich, M.T., and Lewis-Peacock, J.A. (2020). Changes to information in working memory depend on distinct removal operations. Nature Communications *11*, 6239.

Klavir, O., Prigge, M., Sarel, A., Paz, R., and Yizhar, O. (2017). Manipulating fear associations via optogenetic modulation of amygdala inputs to prefrontal cortex. Nature Neuroscience *20*, 836–844.

Klein, A., Ghosh, S.S., Bao, F.S., Giard, J., Häme, Y., Stavsky, E., Lee, N., Rossa, B., Reuter, M., Chaibub Neto, E., et al. (2017). Mindboggling morphometry of human brains. PLoS Computational Biology *13*, e1005350.

Kriegeskorte, N., Goebel, R., and Bandettini, P. (2006). Information-based functional brain mapping. PNAS *103*, 3863–3868.

Lacagnina, A.F., Brockway, E.T., Crovetti, C.R., Shue, F., McCarty, M.J., Sattler, K.P., Lim, S.C., Santos, S.L., Denny, C.A., and Drew, M.R. (2019). Distinct hippocampal engrams control extinction and relapse of fear memory. Nature Neuroscience *22*, 753–761.

Lanczos, C. (1964). Evaluation of Noisy Data. Journal of the Society for Industrial and Applied Mathematics Series B Numerical Analysis *1*, 76–85.

Lawrence, M.A. (2016). ez: Easy Analysis and Visualization of Factorial Experiments.

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

Lenth, R. (2019). Emmeans: estimated marginal means Aka Least-Squares Means. https://cran.r-project.org/package=emmeans.

Letzkus, J.J., Wolff, S.B.E., and Lüthi, A. (2015). Disinhibition, a Circuit Mechanism for Associative Learning and Memory. Neuron *88*, 264–276.

Lissek, S., and van Meurs, B. (2015). Learning models of PTSD: Theoretical accounts and psychobiological evidence. International Journal of Psychophysiology *98*, 594–605.

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

Luke, S.G. (2017). Evaluating significance in linear mixed-effects models in R. Behavior Research Methods *49*, 1494–1502.

Marek, R., Jin, J., Goode, T.D., Giustino, T.F., Wang, Q., Acca, G.M., Holehonnur, R., Ploski, J.E., Fitzgerald, P.J., Lynagh, T., et al. (2018). Hippocampus-driven feed-forward inhibition of the prefrontal cortex mediates relapse of extinguished fear. Nature Neuroscience *21*, 384–392.

Maren, S., Phan, K.L., and Liberzon, I. (2013). The contextual brain: Implications for fear conditioning, extinction and psychopathology. Nature Reviews Neuroscience *14*, 417–428.

Meyer, H.C., Odriozola, P., Cohodes, E.M., Mandell, J.D., Li, A., Yang, R., Hall, B.S., Haberman, J.T., Zacharek, S.J., Liston, C., et al. (2019). Ventral hippocampus interacts with prelimbic cortex during inhibition of threat response via learned safety in both mice and humans. Proceedings of the National Academy of Sciences of the United States of America *116*, 26970–26979.

Milad, M.R., and Quirk, G.J. (2002). Neurons in medial prefrontal cortex signal memory for fear extinction. Nature *420*, 70–74.

Milad, M.R., and Quirk, G.J. (2012). Fear Extinction as a Model for Translational Neuroscience: Ten Years of Progress. Annual Review of Psychology *63*, 129–151.

Milad, M.R., Vidal-Gonzalez, I., and Quirk, G.J. (2004). Electrical stimulation of medial prefrontal cortex reduces conditioned fear in a temporally specific manner. Behav Neurosci *118*, 389–394.

Milad, M.R., Quirk, G.J., Pitman, R.K., Orr, S.P., Fischl, B., and Rauch, S.L. (2007). A Role for the Human Dorsal Anterior Cingulate Cortex in Fear Expression. Biological Psychiatry *62*, 1191–1194.

Moorman, D.E., and Aston-Jones, G. (2015). Prefrontal neurons encode context-based response execution and inhibition in reward seeking and extinction. PNAS.

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

Mumford, J.A., Davis, T., and Poldrack, R.A. (2014). The impact of study design on pattern estimation for single-trial multivariate pattern analysis. NeuroImage *103*, 130–138.

O'Reilly, R.C., and Rudy, J.W. (2001). Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. Psychological Review *108*, 311–345.

Pape, H.C., and Pare, D. (2010). Plastic synaptic networks of the amygdala for the acquisition, expression, and extinction of conditioned fear. Physiological Reviews *90*, 419–463.

Pitman, R.K., Rasmusson, A.M., Koenen, K.C., Shin, L.M., Orr, S.P., Gilbertson, M.W., Milad, M.R., and Liberzon, I. (2012). Biological studies of post-traumatic stress disorder. Nature Reviews Neuroscience *13*, 769–787.

Polyn, S.M., Natu, V.S., Cohen, J.D., and Norman, K.A. (2005). Category-specific cortical activity precedes retrieval during memory search. Science *310*, 1963–1966.

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

Power, J.D., Mitra, A., Laumann, T.O., Snyder, A.Z., Schlaggar, B.L., and Petersen, S.E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. NeuroImage *84*, 320–341.

Qin, C., Bian, X.-L., Wu, H.-Y., Xian, J.-Y., Cai, C.-Y., Lin, Y.-H., Zhou, Y., Kou, X.-L., Chang, L., Luo, C.-X., et al. (2021). Dorsal Hippocampus to Infralimbic Cortex Circuit is Essential for the Recall of Extinction Memory. Cerebral Cortex *31*, 1707–1718.

Raij, T., Nummenmaa, A., Marin, M.-F., Porter, D., Furtak, S., Setsompop, K., and Milad, M.R. (2018). Prefrontal Cortex Stimulation Enhances Fear Extinction Memory in Humans. Biological Psychiatry *84*, 129–137.

Rashid, A.J., Yan, C., Mercaldo, V., Hsiang, H.L., Park, S., Cole, C.J., De Cristofaro, A., Yu, J., Ramakrishnan, C., Lee, S.Y., et al. (2016). Competition between engrams influences fear memory formation and recall. Science *353*, 383–387.

Reijmers, L.G., Perkins, B.L., Matsuo, N., and Mayford, M. (2007). Localization of a Stable Neural Correlate of Associative Memory. Science *317*, 1230–1233.

Ressler, R.L., Goode, T.D., Kim, S., Ramanathan, K.R., and Maren, S. (2021). Covert capture and attenuation of a hippocampus-dependent fear memory. Nat Neurosci.

Ritchey, M., Wing, E.A., LaBar, K.S., and Cabeza, R. (2013). Neural Similarity Between Encoding and Retrieval is Related to Memory Via Hippocampal Interactions. Cerebral Cortex *23*, 2818–2828.

Satterthwaite, T.D., Elliott, M.A., Gerraty, R.T., Ruparel, K., Loughead, J., Calkins, M.E., Eickhoff, S.B., Hakonarson, H., Gur, R.C., Gur, R.E., et al. (2013). An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. NeuroImage *64*, 240–256.

Saygin, Z.M., Kliemann, D., Iglesias, J.E., van der Kouwe, A.J.W., Boyd, E., Reuter, M., Stevens, A., Van Leemput, K., McKee, A., Frosch, M.P., et al. (2017). High-resolution magnetic resonance imaging reveals nuclei of the human amygdala: manual segmentation to automatic atlas. NeuroImage *155*, 370–382.

Seeley, W.W. (2019). The Salience Network: A Neural System for Perceiving and Responding to Homeostatic Demands. J. Neurosci. *39*, 9878–9882.

Senn, V., Wolff, S.B.E., Herry, C., Grenier, F., Ehrlich, I., Gründemann, J., Fadok, J.P., Müller, C., Letzkus, J.J., and Lüthi, A. (2014). Long-range connectivity defines behavioral specificity of amygdala neurons. Neuron *81*, 428–437.

Sierra-Mercado, D., Padilla-Coreano, N., and Quirk, G.J. (2011). Dissociable roles of prelimbic and infralimbic cortices, ventral hippocampus, and basolateral amygdala in the expression and extinction of conditioned fear. Neuropsychopharmacology *36*, 529–538.

Singmann, H., Bolker, B., and Westfall, J. (2015). Analysis of Factorial Experiments, package "afex."

Sotres-Bayon, F., Sierra-Mercado, D., Pardilla-Delgado, E., and Quirk, G.J. (2012). Gating of Fear in Prelimbic Cortex by Hippocampal and Amygdala Inputs. Neuron *76*, 804–812.

Staresina, B.P., Henson, R.N.A., Kriegeskorte, N., and Alink, A. (2012). Episodic reinstatement in the medial temporal lobe. Journal of Neuroscience *32*, 18150–18156.

Staudigl, T., Vollmar, C., Noachtar, S., and Hanslmayr, S. (2015). Temporal-pattern similarity analysis reveals the beneficial and detrimental effects of context reinstatement on human memory. Journal of Neuroscience *35*, 5373–5384.

Talairach, J. (1988). Co-Planar Stereotaxic Atlas of the Human Brain: 3-D Proportional System: An Approach to Cerebral Imaging (Stuttgart ; New York: Thieme).

Taschereau-Dumouchel, V., Cortese, A., Chiba, T., Knotts, J.D., Kawato, M., and Lau, H. (2018). Towards an unconscious neural reinforcement intervention for common fears. Proceedings of the National Academy of Sciences *115*, 3470–3475.

Taschereau-Dumouchel, V., Cortese, A., Lau, H., and Kawato, M. (2020). Conducting decoded neurofeedback studies. Social Cognitive and Affective Neuroscience.

Tovote, P., Fadok, J.P., and Lüthi, A. (2015). Neuronal circuits for fear and anxiety. Nat Rev Neurosci *16*, 317–331.

Tronson, N.C., Schrick, C., Guzman, Y.F., Huh, K.H., Srivastava, D.P., Penzes, P., Guedea, A.L., Gao, C., and Radulovic, J. (2009). Segregated Populations of Hippocampal Principal CA1 Neurons Mediating Conditioning and Extinction of Contextual Fear. J. Neurosci. *29*, 3387–3394.

Tulving, E., and Thomson, D.M. (1973). Encoding specificity and retrieval processes in episodic memory. Psychological Review *80*, 352–373.

Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A., and Gee, J.C. (2010). N4ITK: Improved N3 bias correction. IEEE Transactions on Medical Imaging *29*, 1310–1320.

Twining, R.C., Lepak, K., Kirry, A.J., and Gilmartin, M.R. (2020). Ventral Hippocampal Input to the Prelimbic Cortex Dissociates the Context from the Cue Association in Trace Fear Memory. J. Neurosci. *40*, 3217–3230.

Vallat, R. (2018). Pingouin: statistics in Python. Journal of Open Source Software *3*, 1026.

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

Wheeler, A.L., Teixeira, C.M., Wang, A.H., Xiong, X., Kovacevic, N., Lerch, J.P., McIntosh, A.R., Parkinson, J., and Frankland, P.W. (2013). Identification of a Functional Connectome for Long-Term Fear Memory in Mice. PLOS Computational Biology *9*, e1002853.

Ye, X., Kapeller-Libermann, D., Travaglia, A., Inda, M.C., and Alberini, C.M. (2017). Direct dorsal hippocampal–prelimbic cortex connections strengthen fear memories. Nature Neuroscience *20*, 52–61.

Zhang, X., Kim, J., and Tonegawa, S. (2020). Amygdala Reward Neurons Form and Store Fear Extinction Memory. Neuron *105*, 1077-1093.e7.

Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Transactions on Medical Imaging *20*, 45–57.

SUPPLEMENTARY RESULTS

Interactions in MTL univariate predicting mPFC reinstatement location.

In our analysis using univariate activity at the time of memory retrieval to predict the location of reinstatement in the mPFC, we observed several interactions with hippocampal subfield. For the pHC, there was a significant *pHC * CS type* interaction ($X^2_{(1)}$ = 11.2, P = 8.3e-4), such that the slope of pHC 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 ($X^2_{(1)}$ = 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], P_{FDR} = 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], P_{FDR} = 0.91). There were no significant interactions in the aHC, BLA, or CeM. In sum, MTL univariate activity predicted more reinstatement in the dACC. This effect was stronger for all CS+ items in the pHC compared to CS-, and was selective for conditioning CS+ items in the body of the hippocampus.

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, there was no main effect of *memory accuracy* ($X^2_{(1)}$ = 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* ($X^2_{(1)}$ = 3.08, P = 0.08) and no significant interactions (all Ps ≥ 0.07). In the hippocampus, we again observed a trending main effect of *memory accuracy* ($X^2_{(1)}$ = 3.47, P = 0.06), as well as several significant higher order interactions (*memory accuracy * CS type * encoding context*: $X^2_{(2)}$ = 6.26, P = 0.044; *memory*

*accuracy * CS type * subfield*: $X^2_{(2)} = 6.69$, $P = 0.035$; *memory accuracy * encoding context * group*: $X^2_{(2)} = 11.8$, $P = 0.003$). Thus, recognition memory did not influence reinstatement in the mPFC and amygdala. Recognition memory influencing reinstatement in the hippocampus is consistent with this structure's role in episodic retrieval.

Searchlight Clusters

Group	Encoding context	Label (hemisphere)	MNI coor. peak	Size in voxels ($3mm^3$)
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.