Title: Dissociable neural reinstatement of emotional memories in human PFC

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

**Introduction**

Following a traumatic event, extinction learning allows an organism to form a memory of safety that acts to countervail the original fear association. An adaptive memory system will not normally overwrite threat memories with subsequent experiences of safety, such as during extinction learning. Instead, these opposing associations are stored in a way that allows for the appropriate behavior in response to ambiguous cues in the environment. For example, consider if you were bitten by a snake while hiking in a forest. It would be beneficial to recall this experience in similar settings, such as subsequent hikes, but this memory would not be helpful on your morning commute to work. The neural structures involved in the formation, storage, and retrieval of fear and extinction are increasingly well delineated in the neurobiology of rodents, culminating in the discovery of distinct engrams for each association. There has been some success in translating this work to humans, but several gaps remain. Here, we use a hybrid episodic-associative memory task, combined with multivariate fMRI analyses, to probe the fine-grain organization of fear and extinction memories in the human brain.

During memory retrieval, the neural circuits which originally encoded a memory are reactivated, in a process known as encoding specificity (Tulving & Thomson, 1973). Research seeking evidence of such neural reinstatement has led to great success in determining the location of the exact neural substrates (i.e. engrams) of fear and extinction. Studies of neural activity have long suggested that control of fear and extinction is mediated by a circuit which includes the amygdala, hippocampus, and medial prefrontal cortex (PFC) (Quirk & Mueller, 2008). Recent work has identified encoding specificity of fear and extinction in the amygdala (Grewe et al., 2017; Herry et al., 2008) and hippocampus (Lacagnina et al., 2019). These studies demonstrate that memory storage in the amygdala and hippocampus utilizes a sparse coding scheme, which allows fear and extinction engrams to be stored across different neural populations within the same structures [cite]. Although not spatially segregated, these subcortical ensembles differ in their connectivity with the PFC (Burgos-Robles et al., 2009; Klavir et al., 2017; Marek et al., 2018; Milad & Quirk, 2002; Senn et al., 2014; Sotres-Bayon et al., 2012). In rodents there is a gradient of function along the PFC: the prelimbic (PL) cortex controls the expression of fear, and the infralimbic (IL) cortex controls extinction behavior (Milad & Quirk, 2012). Neural computations across these cortical regions determines the helps determine which association is expressed in response to external cues.

There is converging evidence that a similar circuit for the control fear and extinction exists in humans, although efforts to translate this work using functional MRI neuroimaging techniques have been limited by current methodology. A recent comprehensive fMRI meta-analyses confirmed that the human homologue of the PL, the dorsal anterior cingulate cortex (dACC), is reliably activated during fear learning (Fullana et al., 2016). Maps of whole brain activity during simple discriminatory conditioning tasks were used, screening for brain regions that show differential activity to the CS+ vs. the CS-. However, the meta-analysis failed to detect reliable amygdala or hippocampal activity during fear learning (Fullana et al., 2016). A subsequent meta-analysis of extinction in fMRI also failed to detect amygdala, hippocampus, or ventromedial PFC (vmPFC) activity during extinction learning (Fullana et al., 2018). Thus, it is still unclear whether a similar neural organization exists in the human brain, whereby fear and extinction memories are segregated in separate neural regions and ensembles.

Some success has come from studies utilizing multivariate pattern analysis methods, which are sensitive to the information represented in a pattern of activity instead of average activity over time. Pattern similarity analyses suggest that information about fear and extinction is represented in the amygdala, hippocampus, dACC, and vmPFC (Bach et al., 2011; Graner et al., 2020; Hennings et al., 2020; Visser et al., 2013). We seek to build on this work by providing evidence for neural reinstatement of fear and extinction across the human prefrontal cortex.

Methods to detect neural reinstatement depend on observing the neural activity elicited by the same cue across multiple timepoints. In rodent models, researchers can leverage activity dependent neural tagging methods to observe the reactivation of specific neural ensembles (Lacagnina et al., 2019). Such direct observation remains impossible in humans. However, we show here that multivariate analyses of fMRI can be used to detect evidence of stable and separable memory traces of fear and extinction over time. Our approach relies on the observations that neurocognitive processes active at memory formation are reinstated during retrieval, and that information is often linked to the context in which it was encoded. Previously, *encoding-retrieval similarity* analyses (ERS) have been used to show that during episodic retrieval, patterns of activity corresponding to specific items are reinstated across the human cortex (Johnson et al., 2009; Polyn et al., 2005; Ritchey et al., 2013; Staresina et al., 2012; Staudigl et al., 2015). Other studies have shown that retrieval of an episodic or associative memory is accompanied by the neural reinstatement of specific mental context in which the memories were encoded (Bornstein & Norman, 2017; Gershman et al., 2013; Hennings et al., 2020; Manning et al., 2016). We used these properties of episodic memory to probe the reinstatement of fear and extinction in humans, in a way similar to the principle behind activity dependent neural tagging in rodents.

We have previously reported on our category conditioning task (Dunsmoor & Kroes, 2019), in which semantic categories (e.g. animals and tools) serve as conditioned stimuli. A key feature of this task is that while participants are undergoing fear conditioning and extinction, they are also forming an episodic memory for each unique category exemplar. After a test of fear renewal the next day, participants then underwent a surprise recognition memory test for the images they saw during conditioning and extinction the previous day. Crucially, both associative learning and the recognition memory test was completed during fMRI, allowing us to use an encoding-retrieval similarity analysis to probe for neural reinstatement. For each image, the pattern of activity elicited during encoding was correlated with the pattern elicited during the retrieval test. We then tested if neural reinstatement of these memories varied in our *a priori* ROIs based on their emotional association (CS+ or CS-) or the context in which they were encoded (fear or extinction). Based on previous work in both rodents and humans, we predicted that we would observe neural reinstatement of fear in the dACC, and extinction reinstatement in the vmPFC. As well as probing reinstatement in the hippocampus and amygdala, we compare neural signatures of memory fidelity between healthy participants and individuals with post-traumatic stress symptoms (PTSS). PTSS is characterized by both over-expression of fear, and decreased extinction retrieval. One possible explanation for these behavioral symptoms could dysregulated organization of these emotional associations in the brain. We hope that linking multivariate signatures of fear and extinction to the pathophysiology of PTSD will have direct benefit to the its treatment.

**RESULTS**

**Chart, line chart

Description automatically generated**

**Figure 1. Encoding-retrieval similarity in the PFC.** Healthy adults exhibited a double dissociation of emotional reinstatement (CS+ - CS-) in the PFC, such that fear reinstatement was higher in the dACC, and extinction reinstatement was higher in the vmPFC. In PTSS, the dACC displayed significant emotional reinstatement of extinction memories, representing a misallocation. Error bars correspond to the 95% confidence interval of the CS+/- difference. \*\*\*P < 0.001, \*\*P < 0.01, \*P < 0.05.

**Encoding-retrieval similarity in the PFC**

Trial-wise encoding-retrieval similarity across the PFC was modeled using a linear mixed effects approach, with fixed effects of CS condition (CS+/-), encoding phase (pre-conditioning, fear conditioning, or extinction), ROI (dACC or vmPFC), and group (healthy or PTSS), and included a random intercept of subject (*ERS* ~ *CS* *condition\*encode phase\*ROI\*group* + (1|*subject*); note that main effects are implicitly coded within the interaction term). Likelihood ratio tests were used to test the main effects and interactions of fixed effects. Overall participants exhibited significant positive ERS in the PFC, as evidenced by a significant intercept (0.119, 95% CI = [0.094, 0.144], P = 1.44e-20). We first examined main effects, and found that ERS in the PFC was sensitive to both the emotional association (*CS condition* (Chisq (1) = 46.5, P = 8.35e-12) and encoding phase of each item (*encode phase* (Chisq (2) = 49.5, P = 1.77e-11). On the other hand, total ERS did not differ between the dACC and vmPFC (*ROI* (Chisq (1) = 2.50, P = 0.11), and there was no overall difference in ERS between Healthy and PTSS participants (*group* Chisq (1) = 0.043, P = 0.83). We found that *CS condition* interacted significantly with several other factors, including *condition*\**encode phase* (Chisq (2) = 44.7, P = 1.99e-10), *condition*\**ROI* (Chisq (1) = 12.2, P = 4.62e-4) interactions, as well as the three-way interaction *condition*\**encode phase*\**ROI* (Chisq (2) = 12.5, P = 0.002). These interactions support the hypothesis that different regions of PFC exhibit preferential neural reinstatement of items based on their emotional content and encoding context. We also found a significant *encode phase\*group* interaction (Chisq (2) = 7.01, P = 0.030), as well as a three-way interaction *condition*\**ROI*\**group* (Chisq (1) = 4.20, P = 0.040). These *group* interactions support the hypothesis that prior exposure to trauma (PTSS) affects how associative emotional memories are reinstated in the PFC.

Based on both our initial hypotheses and the pattern of main effects and interactions, we conducted follow up tests focusing on the effects of *CS condition.* Specifically, for each encoding phase, ROI, and group we tested the difference in ERS between the CS+ and CS-, which we refer to as emotional reinstatement (Figure 1 or 2). These comparisons were accomplished using tests of marginal means of the linear mixed effects model, and false discovery rate (FDR) corrections were applied using the *p.adjust* function in R (12 tests total). Healthy individuals exhibited significant emotional reinstatement for items encoded during fear conditioning (CS diff. = 0.222, [0.163, 0.283], PFDR = 4.62e-12), consistent with a for the dACC in both threat acquisition and retrieval. In the vmPFC, healthy individuals exhibited significant emotional reinstatement for items encoded during both fear conditioning (CS diff. = 0.074, [0.013, 0.134], PFDR = 0.033) and extinction (CS diff. = 0.113, [0.053, 0.173], PFDR = 9.20e-4). These results suggest that in healthy adults the vmPFC exhibits encoding specificity of both fear and extinction memories, which is not detected in normal univariate fMRI analyses. Like their healthy counterparts, individuals with PTSS exhibited significant emotional reinstatement in the dACC for items encoded during fear conditioning (CS diff. = 0.171, [0.111, 0.231], PFDR = 1.53e-7). However unlike in healthy participants, individuals with PTSD also showed significant emotional reinstatement of items encoded during extinction in the dACC (CS diff. = 0.103, [0.043, 0.164], PFDR = 0.002). In contrast, in the vmPFC PTSS individuals did not show significant emotional reinstatement for either fear conditioning or extinction. These results suggest that the organization of associative memories in the PFC is dysregulated in individuals PTSS; information about safety learning is being misallocated to the dACC, a region that normally only codes for memories of threat. PTSS individuals did exhibit negative emotional reinstatement in the vmPFC for items encoded during pre-conditioning (CS diff. = -0.079, [-0.139, -0.019], PFDR = 0.024). Although not directly related to our hypotheses, this difference could be the result of the general vmPFC dysfunction exhibited in this group.

These results establish that there is significant emotional neural reinstatement in both dACC and vmPFC. We next performed follow up contrasts to test for a double dissociation of emotional reinstatement across these two regions. Specifically, we took double subtraction of the CS+ - CS- for either conditioning or extinction between the dACC and vmPFC. FDR correction was applied across the 4 tests. For healthy adults, the dACC exhibited significantly more emotional reinstatement for fear conditioning memories as compared to the vmPFC (ROI diff. = 0.149, [0.064, 0.234], PFDR = 0.002). The converse was also true; that is, the vmPFC showed significantly more emotional reinstatement for extinction memories as compared to the dACC (ROI diff. = 0.099, [0.014, 0.184], PFDR = 0.031). Thus in healthy adults, the PFC exhibits a double dissociation in the neural reinstatement of emotional memories, based on the valence of associative learning during memory encoding. Similar to the healthy adults, individuals with PTSS show increased emotional reinstatement for fear conditioning memories in the dACC compared to the vmPFC (ROI diff. = 0.121, [0.036, 0.206], PFDR = 0.011). However, there was no difference between the two regions for the emotional reinstatement of items encoded during extinction in PTSS (ROI diff = -0.063, [-0.148, 0.023], PFDR = 0.15).

**Encoding-retrieval similarity in subcortical ROIs**

**Chart

Description automatically generated**

**Figure 2. Hippocampal ERS.** Healthy adults exhibited a double dissociation in the reinstatement of fear and extinction between the anterior and posterior hippocampus, with more extinction reinstatement in the anterior, and more fear reinstatement in the posterior. In PTSS the posterior hippocampus was sensitive to fear reinstatement, but there was no significant reinstatement of extinction memories in the anterior. No significant pairwise differences by encoding phase were observed in the body of the hippocampus (data not shown). Data are averaged across CS+/- for each phase, see text for details. Error bars correspond to the 95% confidence interval of each marginal mean. \*\*\*P < 0.001, \*P < 0.05.

*Hippocampal ERS*. The amygdala and hippocampus are crucial structures for the acquisition and expression of both learned fear and extinction. The hippocampus in particular exerts contextual control over fear and extinction. When a previously extinguished stimulus is encountered in a novel context, the hippocampus suppresses the expression of extinction in favor of fear. Research suggests that different portions of the hippocampus along its long axis serve different functions in the course of conditioning and extinction, however there is not a clear pattern. For example the anterior hippocampus (ventral in rodents) has been shown to drive both the inhibition of fear (Meyer et al., 2019) and fear relapse (Marek et al., 2018), seeming to depend on connections to the PFC. Similarly, the posterior hippocampus (dorsal in rodents) has been shown to be necessary for acquisition of fear (Bast et al., 2003) and extinction (Corcoran et al., 2005) in contextual conditioning designs.

Using the automatic hippocampal segmentation provided in Freesurfer, we probed ERS along the long axis of the hippocampus. Three bi-lateral subfields were used: head (anterior), body, and tail (posterior). As in our analysis of ERS in the PFC, we entered trial-wise data into linear mixed effect model (*ERS* ~ *CS condition*\**encode phase*\**ROI*\**group* + (1|*subject*), where *ROI* here refers to hippocampal subfield. The model intercept was significantly positive (0.021, [0.012, 0.030], P = 2.89e-6), indicating that overall there was significant ERS in the hippocampus. However, there were no significant main effects in the model (*CS condition* Chisq (1) = 0.135, P = 0.71; *encode phase* Chisq (2) = 3.21, P = 0.20; *ROI* Chisq (2) = 1.26, P = 0.53; *group* Chisq (1) = 0.213, P = 0.64). Despite the lack of main effects, there was a significant *encode phase*\**ROI* interaction (Chisq (4) = 22.95, P = 1.29e-4) and a significant three-way interaction of *encode phase*\**ROI*\**group* (Chisq (4) = 12.82, P = 0.012). These interactions suggest that ERS along the long axis is sensitive to encoding phase, or context, and this may be further modulated by previous exposure to trauma (PTSS).

We further explored these interactions with pairwise comparisons of *encode phase*, within each of the three subfields and group (FDR correction applied across the 18 tests). Since *CS condition* did not significantly contribute to any effects or interactions, comparisons of encoding phase were averaged across CS+/-. In healthy adults, we found that posterior hippocampus exhibited significantly more ERS for items encoded during conditioning compared to extinction (phase diff. = 0.044, [0.014, 0.073], PFDR = 0.019). In contrast, the anterior hippocampus displayed the opposite effect, that is ERS was higher for extinction compared to conditioning (phase diff. = 0.065, [0.035, 0.094], PFDR = 3.36e-4). There were no significant differences in the body of the hippocampus for healthy adults. These results support the idea that the hippocampus reinstates contextual information of fear and extinction, and that this reinstatement differs along the long axis. As with healthy adults, we found that in PTSS the posterior hippocampus showed selective ERS for the conditioning context, significantly greater than both pre-conditioning (phase diff. = 0.044, [0.014, 0.074], PFDR = 0.019) and extinction (phase diff. = 0.043, [0.014, 0.073], PFDR = 0.019). Unlike in the healthy group, PTSS showed no significant differences in the anterior hippocampus. The lack of selective extinction ERS in the anterior hippocampus is more evidence that the organization of safety memories is dysregulated in individuals with PTSS. Again, there were also no significant differences between phases in the body of the hippocampus.

Given this pattern of results within the anterior and posterior hippocampus, we next explored whether there was a significant double dissociation of extinction and fear between these subdivisions of the hippocampus (FDR correction across 4 tests). In healthy adults, there was indeed a double dissociation, such that ERS of extinction was higher in the anterior hippocampus compared to the posterior (ROI diff. = 0.075, [0.046, 0.105], PFDR = 2.51e-6), and conditioning was higher in the posterior compared to the anterior (ROI diff. = 0.033, [0.003, 0.063], PFDR = 0.038). In individuals with PTSS, we did not observe any difference between anterior and posterior hippocampus for extinction items (ROI diff. = -0.004, [-0.034, 0.026], PFDR = 0.80), which is not surprising given the similarly low amounts of extinction related ERS in both regions. As in the healthy group, the posterior hippocampus displayed significantly more ERS for conditioning items compared to the anterior hippocampus in PTSS (ROI diff. = 0.034, [0.004, 0.063], PFDR = 0.038). In addition to the vmPFC, these results suggest that the anterior hippocampus is also dysregulated when it comes to extinction related processing in PTSS.

*Amygdala ERS*. We next examined ERS in the amygdala. The amygdala was segmented into two bilateral ROIs based on their roles in the acquisition and expression of fear and extinction, namely the basolateral amygdala (BLA) and the central nucleus of the amygdala (CeM). We again entered trial-wise data into linear mixed effect model (*ERS* ~ *CS condition*\**encode phase*\**ROI*\**group* + (1|*subject*), where *ROI* here refers to amygdalar subfield. There was overall significant ERS in the amygdala (intercept = 0.012, [0.003, 0.022], P = 0.008), however we note that intercept of this model is qualitatively lower than in the hippocampus or PFC. The model contained no significant main effects (*CS condition* Chisq (1) = 0.001, P = 0.98; *encode phase* Chisq (2) = 3.35, P = 0.19; *ROI* Chisq (1) = 1.47, P = 0.23; *group* Chisq (1) = 0.188, P = 0.66), and surprisingly there were also no significant interactions. As such, we did not perform post-hoc tests, and explore reasons for why amygdala ERS did not vary amongst experimental variables in the discussion.

**Subcortical neural signatures predict differential ERS in the PFC**

In healthy adults, we observed a double dissociation of ERS in the PFC, with fear memories being preferentially reinstated in the dACC, and extinction memories being preferentially reinstated in the vmPFC. We next asked if other neural signatures could predict the dissociation observed in the PFC. Specifically, we tested in turn whether subcortical univariate activity during retrieval, subcortical ERS, or a signature of extinction context reinstatement predicted the trial-by-trial difference in PFC encoding-retrieval similarity. Since our goal was to predict the balance in ERS between the vmPFC and dACC, for each trial we took the difference in ERS between these two regions (vmPFC – dACC) as the outcome variable of interest. We restricted our analysis to items encoded during conditioning and extinction as our encoding contexts of interest. As discussed below, we tested whether various neural signatures could predict the difference in PFC neural reinstatement on each trial using linear mixed effects models. The form of the model was thus: *PFC difference* ~ *predictor*\**CS condition*\**encode phase*\**group* + (1|*subject*), and our analysis focused on testing main effects and interactions with *predictor* for each of the different neural signatures tested. The directionality of main effects was determined by the fixed effect slope, and the directionality of interactions by testing the marginal slopes. In this model, a significant positive slope indicates the a predictor signals more reinstatement in the vmPFC, while a significant negative slope indicates the predictor signals more reinstatement in the dACC.



**Figure 3. Factors predicting differential reinstatement in the PFC**. We tested, in turn, whether univariate activity or encoding retrieval similarity in each of our subcortical ROIs could predict the trial-by-trial difference in reinstatement in the vmPFC vs. dACC. Univariate activity in all ROIs predicted a bias in reinstatement to the dACC, while only ERS in the anterior hippocampus predicted a bias to the vmPFC. Slope estimates for ERS and univariate activity varied drastically in their magnitude, and are shown zoomed in to facilitate qualitative comparisons within modality. Error bars correspond to the 95% confidence interval of the overall slope of the predictor in each model. Significance was determined using likelihood ratio tests. \*\*\*P < 0.001, \*\*P < 0.01, \*P < 0.05.

*Subcortical univariate activity*. We first tested whether univariate activity in the subfields of the hippocampus or amygdala during the retrieval test predicted the split in PFC reinstatement for each item. We found that all subfields of both structures were significant negative predictors, suggesting that more subcortical activity supports more ERS in the dACC (posterior hippocampus: Chisq (1) = 54.7, P = 1.38e-13, slope = -1.8e-3; hippocampus body: Chisq (1) = 68.2, P = 1.48e-16, slope = -2.49e-3; anterior hippocampus: Chisq (1) = 46.8, P = 8.00e-12, slope = -1.7e-3; BLA: Chisq (1) = 26.7, P = 2.39e-7, slope = -1.45e-3; CeM: Chisq (1) = 19.5, P = 1.01e-5, slope = -6.90e-4). Additionally, we observed several interactions with hippocampal subfields. In the posterior hippocampus model, there was a significant *posterior HC*\**CS* *condition* (Chisq (1) = 11.2, P = 8.3e-4), such that the slope of posterior hippocampal activity was significantly more negative for CS+ compared to CS- (CS slope diff. = -1.53e-3, [-2.43e-3, -6.34e-4], P = 8.24e-4). In the hippocampus body model, there was a significant *HC body*\*CS condition (Chisq (1) = 6.25, P = 0.012) as well as a *HC body*\**CS condition*\**encode phase* interaction (Chisq (1) = 5.46, P = 0.019). To explore this interaction, we tested the CS+ - CS- difference in slope for each phase (FDR correction for 2 tests). We found 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 interaction 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, and selectively stronger for CS+ from acquisition in the body of the hippocampus.

*Subcortical ERS*. Next, we assessed whether the amount of neural reinstatement in these subcortical regions predicted where in the PFC each item was preferentially reinstated. It should be noted that even though univariate activity in these regions all predicted more reinstatement in the dACC, this is not automatically the hypothesis for what ERS will predict. The information represented in a pattern of activity, captured by ERS, is a different signal from the total average activation across that pattern.

The posterior hippocampus was a significant negative predictor (Chisq (1) = 4.64, P = 0.031, slope = -0.060), indicating that more reinstatement in the posterior hippocampus is associated with more reinstatement in the dACC. There were no significant interactions with the other terms in the model. Even though there were no interactions, we performed an exploratory analysis to specifically test the significance of the marginal slopes within encoding phase and group (FDR correction across 4 tests against 0). This analysis was motivated by the strong bias in fear over extinction reinstatement we observed in both groups in the posterior hippocampus. Interestingly, in healthy adults the slope was not significant for either conditioning (slope = 0.052, [-0.062, 0.167], PFDR = 0.40) or extinction (slope = -0.087, [-0.201, 0.027], PFDR = 0.40). In PTSS, the slope for conditioning was initially significant (P = 0.020) but did not survive FDR correction (slope = -0.128, [-0.233, -0.020], PFDR = 0.079). The slope for extinction was also not significant in PTSS (slope = -0.080, [-0.183, -0.024], PFDR = 0.40). As a final test, we directly compared the slope of conditioning between the two groups, and found that it was significantly more negative for PTSS compared to healthy (group slope diff. = 0.179, [0.023, 0.335], P = 0.025). The main effect observed in the posterior hippocampus thus seems to be primarily driven by a select relationship within the subset of neural reinstatement of fear conditioning in the PTSS group.

Reinstatement in body of the hippocampus was not a significant predictor (Chisq (1) = 1.37, P = 0.24), nor were there any significant interactions. In contrast to the posterior region, reinstatement in the anterior hippocampus was a significant positive predictor (Chisq (1) = 11.1, P = 8.51e-4, slope = 0.091). The positive slope indicates that neural reinstatement in the anterior hippocampus is associated with more reinstatement in the vmPFC compared to the dACC. As with the posterior hippocampus, there were no significant interactions with anterior hippocampal ERS. However, we again conducted exploratory analyses of the marginal slopes within phase and group (FDR correction across 4 tests against 0). Previously, we had shown that the anterior hippocampus exhibited a selective extinction vs. conditioning bias in reinstatement, but only in healthy adults. As such, we wanted to probe the significant main effect observed here along those same dimensions. Indeed, we found that in healthy adults, the relationship for conditioning was not significant (slope = 0.059, [-0.037, 0.156], PFDR = 0.68), but was significant for extinction (slope = 0.214, [0.104, 0.323], P­FDR = 5.48e-4). In PTSS, the relationship was not significant for either conditioning (slope = 0.055, [-0.050, 0.156], PFDR = 0.68) or extinction (slope = 0.037, [-0.078, -0.152], PFDR = 0.68). We then directly compared the slopes for extinction between the two groups, and found that the relationship was significantly stronger in healthy adults compared to PTSS (group slope diff. = 0.176, [0.017, 0.335], P = 0.030). The main effect observed in the anterior hippocampus is being driven by a select relationship within the subset of neural reinstatement of extinction in the healthy group.

We next assessed whether ERS in the amygdala could predict the split in prefrontal reinstatement. Reinstatement in the BLA was not a significant predictor (Chisq (1) = 0.404, P = 0.52). Surprisingly, there were significant *BLA*\**condition*\**group* (Chisq (1) = 4.45, P = 0.035) and *BLA*\**phase*\**group* interactions (Chisq (1) = 4.48, P = 0.034). However, we did not conduct post-hoc analyses given the lack of both a main effect and no variability in BLA reinstatement. The CeM however was a significant negative predictor (Chisq (1) = 8.49, P = 0.004, slope = -0.065). There were no significant interactions with CeM reinstatement, and we did conduct any exploratory analyses. In sum, reinstatement in the CeM predicted a bias in prefrontal reinstatement to the dACC.

*Separable influence of the anterior hippocampus.* In our analyses, we found that different neural signals from the anterior hippocampus significantly predict a bias in reinstatement in opposite directions. Anterior hippocampal univariate activity predicted more reinstatement in the dACC, while ERS in the same region predicted more reinstatement in the vmPFC. A region which exhibits bi-directional control could be crucial for the proper regulation of fear and extinction in humans. As such, we tested whether these effects were separable in a single model: *PFC difference* ~ *anterior HC ERS*\**anterior HC univariate*\**CS condition*\**phase*\**group* + (1|*subject*). Indeed, in this model anterior hippocampal ERS was again a significant positive predictor (Chisq (1) = 5.56, P = 0.018, slope = 0.067), and univariate activity in the same region was at the same time a significant negative predictor (Chisq (1) = 42.5, P = 7.1e-11, slope = -1.65e-3). There were no significant interactions with either predictor. These results suggest that in the anterior hippocampus, more total activity predicts a bias in reinstatement to the dACC, while more local reinstatement predicts a bias in reinstatement to the vmPFC.

Chart, scatter chart

Description automatically generated

**Figure 4. Reinstated extinction context predicts ERS bias to the vmPFC for CS+ items only.** Estimates of the marginal slope for each condition are shown with 95% CI. Reinstated extinction context was a significant positive predictor for CS+ items (P = 1.84e-4), but not for CS- (P = 0.18). Participant-averaged data are shown for visualization purposes, although the model was fit with trial-wise data. Data are averaged across encoding phase.

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

A feature of the experimental design was that during extinction learning on day 1, trial irrelevant “context tags” were inserted between each trial. Previously, these method has been used to track the reinstatement of a specific episodic mental context during memory retrieval (Bornstein & Norman, 2017; Gershman et al., 2013; Manning et al., 2016). We previously reported how during the test of fear renewal on day 2, the context tags can be used to track the reinstatement of the previously encoded extinction context (Hennings et al., 2020). We showed that this neural marker of extinction context was a proxy for the associative memory of extinction: in healthy adults the degree of extinction context reinstatement predicted the behavioral expression of extinction and activity in the vmPFC. Since this extinction context tag was already present in the experimental design, we explored whether evidence for the reinstatement of the extinction context during the recognition memory test predicted the bias in reinstatement in the PFC. Based on our previous work, we predicted that if there was an effect, it would be to the vmPFC.

The extinction mental context tag was built by showing natural scene images during the it is of extinction learning. During later timepoints, when natural scene images are not shown, reinstatement of the extinction context is inferred by the amount of classifier evidence for scenes in the parahippocampal place area (PPA). Data from the perceptual localizer task collected at the end of day 2 is used to functionally define the PPA, and to train the MVPA classifier (see Methods for more details). We then used the classifier to predict the amount of scene-related information in the PPA, which again is a proxy for reinstatement of the extinction context. The trial-wise estimates were then used to predict the differential split in ERS observed in the PFC, similar to the subcortical neural signatures tested above (*PFC difference* ~ *extinction context*\**CS condition*\**encode phase*\**group +* (1|*subject*)). We again excluded trials from pre-conditioning to focus on our effects of interest in the PFC. In this model extinction context evidence was a significant positive predictor (Chisq (1) = 12.6, P = 3.92e-4, slope = 0.093). In addition, there was a trending *extinction context*\**CS condition* interaction (Chisq (1) = 3.66, P = 0.056). We explored this interaction by testing the marginal slope of *extinction context* in each CS, collapsed across phase and group (FDR correction for 2 tests against 0). Reinstated extinction context was a significant positive predictor for CS+ (slope = 0.139, [0.069, 0.208], PFDR = 1.84e-4), but was not significant for CS- (slope = 0.048, [-0.021, 0.117], PFDR = 0.18). These results indicate for emotional items (CS+), the degree of extinction context reinstatement predicts the bias in prefrontal reinstatement to the vmPFC.