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

**Abstract**

Specifically, we sought to isolate quantifiable and separate memory traces of fear and extinction in the human brain.

**Introduction**

Following a traumatic event, extinction learning allows an organism to form a memory of safety that acts to countervail the original fear association. However, an adaptive memory system will not overwrite threat memories with subsequent experiences of safety. Instead, these opposing associations are stored in a way that allows the appropriate 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 has 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 separated 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 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**

**Encoding-retrieval similarity in the PFC**

Encoding-retrieval similarity across the PFC was modeled using a linear mixed effects model, with fixed effects of CS condition, encoding phase, ROI, and group, and included a random intercept of subject (ers ~ condition\*phase\*roi\*group + (1|subject)). Likelihood ratio tests were used to test the main effects and interactions of fixed effects, as well as the random intercept. See Methods for more details.

Overall, participants exhibited significant ERS in the PFC, as evidenced by a significant intercept Chisq = . In addition, there were significant main effects of both *CS condition* (Chisq (1) = 46.51, P = 8.35e-1) and *encoding phase* (Chisq (2) = 49.51, P = 1.77e-11).

In the dACC, main effect of phase F(2, 6743.2) = 19.81, P = 2.64e-9, condition F(1,6743.2) = 57.59, P = 3.67e-14, a phase by condition interaction F(2, 6743.2) = 19.78, P = 2.75e-9, and a marginally significant phase by condition by group three way interaction F(2, 6743.2) = 2.85, P = 0.0579). Planned post-hoc comparisons included CS+ vs. CS- within each experimental phase for each group. For the healthy adults: post-hoc stuff here

In the vmPFC, main effect of phase F(

**METHODS**

Anatomical data preprocessing

The T1-weighted (T1w) image was corrected for intensity non-uniformity (INU) with N4BiasFieldCorrection (Tustison et al., 2010), distributed with ANTs 2.2.0 (Avants et al., 2008), and used as T1w-reference throughout the workflow. The T1w-reference was then skull-stripped with a *Nipype* implementation of the antsBrainExtraction.sh workflow (from ANTs), using OASIS30ANTs as target template. Brain tissue segmentation of cerebrospinal fluid (CSF), white-matter (WM) and gray-matter (GM) was performed on the brain-extracted T1w using fast (FSL 5.0.9, Zhang et al., 2001). Brain surfaces were reconstructed using recon-all (FreeSurfer 6.0.1, Dale et al., 1999), and the brain mask estimated previously was refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray-matter of Mindboggle (Klein et al., 2017). Volume-based spatial normalization to one standard space (MNI152NLin2009cAsym) was performed through nonlinear registration with antsRegistration (ANTs 2.2.0), using brain-extracted versions of both T1w reference and the T1w template. The following template was selected for spatial normalization: *ICBM 152 Nonlinear Asymmetrical template version 2009c* (Fonov et al., 2009).

Functional data preprocessing

For each of the 9 BOLD runs found per subject (across all tasks and sessions), the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. Susceptibility distortion correction (SDC) was omitted as no field maps were collected. The BOLD reference was then co-registered to the T1w reference using bbregister (FreeSurfer) which implements boundary-based registration (Greve & 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 & Hyde, 1997). The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying the transforms to correct for head-motion. These resampled BOLD time-series will be referred to as *preprocessed BOLD in original space*, or just *preprocessed BOLD*. The BOLD time-series were resampled into standard space, generating a *preprocessed BOLD run in MNI152NLin2009cAsym space*. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. Several confounding time-series were calculated based on the *preprocessed BOLD*: framewise displacement (FD), DVARS and three region-wise global signals. FD and DVARS are calculated for each functional run, both using their implementations in *Nipype* (following the definitions by (Power et al., 2014). The three global signals are extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (*CompCor*, Behzadi et al., 2007). Principal components are estimated after high-pass filtering the *preprocessed BOLD* time-series (using a discrete cosine filter with 128s cut-off) for the two *CompCor* variants: temporal (tCompCor) and anatomical (aCompCor). tCompCor components are then calculated from the top 5% variable voxels within a mask covering the subcortical regions. This subcortical mask is obtained by heavily eroding the brain mask, which ensures it does not include cortical GM regions. For aCompCor, components are calculated within the intersection of the aforementioned mask and the union of CSF and WM masks calculated in T1w space, after their projection to the native space of each functional run (using the inverse BOLD-to-T1w transformation). Components are also calculated separately within the WM and CSF masks. For each CompCor decomposition, the *k* components with the largest singular values are retained, such that the retained components’ time series are sufficient to explain 50 percent of variance across the nuisance mask (CSF, WM, combined, or temporal). The remaining components are dropped from consideration. The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. The confound time series derived from head motion estimates and global signals were expanded with the inclusion of temporal derivatives and quadratic terms for each (Satterthwaite et al., 2013). Frames that exceeded a threshold of 0.5 mm FD or 1.5 standardized DVARS were annotated as motion outliers. All resamplings can be performed with *a single interpolation step* by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and output spaces). Gridded (volumetric) resamplings were performed using antsApplyTransforms (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels (Lanczos, 1964). Non-gridded (surface) resamplings were performed using mri\_vol2surf (FreeSurfer).

**Region of interest selection**

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

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). The hippocampus was segmented into head (anterior), body, and tail (posterior) subfields along the long axis. Anterior to posterior segmentation was used based on previous research which suggests different roles in fear and extinction for the homologous structures in the rodent hippocampus [Cite]. The amygdala was segmented into the basolateral (BLA), and central nucleus (CeM) subfields based on their functional roles in associative learning and memory expression [good cite]. The anatomical segmentations were registered to functional space using mri\_label2vol and binary masks created using fslmaths.

A functionally defined parahippocampal place area (PPA) mask was created to facilitate decoding of the mental context tag learned during extinction (Bornstein & Norman, 2017; Hennings et al., 2020). See Hennings et al., (2020) for a full description of the localizer task used. In brief, a GLM was used to estimate the contrast Scenes > Scrambled Scenes and Objects. Each subject’s estimate was threshold at p=0.001 uncorrected, binarized, and stacked. A cluster was corresponding to the PPA was selected using the criteria that at least 95% of subjects show activation. The mask was registered to functional resolution using flirt with 12 degrees of freedom and nearest neighbor interpolation for each subject.

**Multivariate pattern analysis**

*Neural decoding.* Extinction mental context tags were decoded using the same procedure as in Hennings et al., (2020), implemented in Python. In brief, a logistic regression classifier (Sklearn, (Pedregosa et al., 2012) was trained to discriminate natural scene images vs. scrambled images [check]…

*Encoding-retrieval similarity analysis*. In