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. Since Pavlov, we have known that extinction is new learning, and does not overwrite the original threat association. How are these opposing associations of threat and safety organized in the brain? An adaptive memory system should not overwrite threat memories given an experience of safety, instead storing them in a way that allows an organism to respond appropriately to ambiguous cues in the environment that may signal either threat or safety. This normally adaptive organization of emotional associations may be dysregulated in psychiatric disorders such as post-traumatic stress disorder (PTSD), which is characterized by excessive fear expression and impaired extinction memory retrieval. The neural mechanisms involved in the formation, storage, and retrieval of fear and extinction are increasingly well delineated in the neurobiology in rodents, culminating in the discovery of distinct engrams for each association. Translating this work to humans has lagged using simple discriminatory conditioning paradigms and traditional fMRI analysis techniques [for example, lack of amygdala and vmPFC]. Here, we used a multivariate approach of fMRI data combined with a hybrid episodic-associative memory task to probe the fine-grain organization of fear and extinction memories in the human brain.

During memory retrieval, the neural circuits which encoded an experience are reactivated, in a process known as encoding specificity. Just as memory retrieval strength is a function of the degree to which cues, external or internal, between encoding and retrieval match, the same is true of neural ensembles. The higher the degree of overlap in activity across the circuit leads to better memory retrieval (as evidenced by behavior). This process has been demonstrated in both humans and rodent models.

This work relies on our ability to tag and track neural activity at multiple timepoints. Fear learning and fear extinction are ideal behaviors in which to probe for engram like activity in the brain, as the neural circuits involved are well mapped, and the behaviors are expressed specific patterns (e.g., they are contextually specific) (Bouton, 2002). Neurophysiological research in rodents shows distinct representations of fear versus extinction memory traces coded withing and between the amygdala, hippocampus, and medial prefrontal cortex (PFC) (Quirk & Mueller, 2008). Previous work has identified engrams corresponding to fear and extinction in the amygdala (Grewe et al., 2017; Herry et al., 2008) and hippocampus (Lacagnina et al., 2019). Memory storage in the amygdala and hippocampus relies on a sparse coding scheme, which allows fear and extinction engrams to be stored across different neural circuits within the same structures [cite]. Although not spatially segregated, fear and extinction ensembles differ in their long range connections and interactions with the PFC (Burgos-Robles et al., 2009; Marek et al., 2018; Milad & Quirk, 2002; Sotres-Bayon et al., 2012). In the rodent PFC, the prelimbic (PL) cortex controls the expression of fear, and the infralimbic (IL) cortex controls extinction behavior (Milad & Quirk, 2012).

There is converging evidence that a similar circuit for the control fear and extinction exists in humans, although progress in translating the work from rodent models has perhaps been limited by current methodology., Recent comprehensive fMRI meta-analyses investigated whole brain activity during simple discriminatory conditioning tasks, screening for brain regions that show differential activity to the CS+ vs. the CS-. The human homologue of the PL, the dorsal anterior cingulate cortex (dACC) is reliably activated during fear A studies failed to detect reliable amygdala activity during fear conditioning (Fullana et al., 2016), and failed to detect amygdala or vmPFC activity during extinction learning (Fullana et al., 2018). In sum, is insufficient to elucidate the circuit.

Studies utilizing context conditioning paradigms have confirmed a role of the hippocampus in exerting contextual control over extinction

separation of function exists in the human homologues of these structures, the dorsal anterior cingulate cortex (dACC) and the ventral medial PFC (vmPFC). Specifically, neuroimaging research utilizing whole brain activations reveal a role of the dACC in the acquisition of conditioned fear. Whole brain activations have not consistently revealed the vmPFC as playing a crucial role in fear extinction processes in humans; however this is most likely a feature of the standard method used to detect differential univariate activity in fMRI studies of fear and extinction. Indeed, other contrast methods and information sensitive multivariate pattern analysis (MVPA) techniques suggest that the vmPFC is involved in extinction processes. However, unlike in work in rodent models, human research has been unsuccessful in finding evidence of encoding specificity of these emotional associations in the PFC.

In humans, this idea has been formalized into computational models, such as the context maintenance and retrieval model (CMR), which posits that contextual representations (mapped onto patters of neural activity) are reinstated during retrieval processes. Recently, this has been extended to explain the retrieval of emotional memories as well.

Here we sought to isolate quantifiable and separate memory traces of fear and extinction in the human brain, to discover whether and how memory representations of extinction change over time. Additionally, we compared neural signatures of extinction memory fidelity between healthy adults and individuals with PTSD.

**RESULTS**

**Encoding-retrieval similarity in the PFC**

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).