**ONLINE METHODS**

**Participants***.*

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

**Stimuli.**

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

**Task.**

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

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

**Functional MRI acquisition.**

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

**Image preprocessing**

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

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

*Functional data preprocessing.* For each of the 9 BOLD runs found per subject (across all tasks and sessions), the following preprocessing was performed. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. Susceptibility distortion correction (SDC) was omitted as no field maps were collected. The BOLD reference was then co-registered to the T1w reference using bbregister (FreeSurfer) which implements boundary-based registration (Greve and Fischl, 2009). Co-registration was configured with six degrees of freedom. Head-motion parameters with respect to the BOLD reference (transformation matrices, and six corresponding rotation and translation parameters) are estimated before any spatiotemporal filtering using mcflirt (FSL 5.0.9, (Jenkinson et al., 2002). BOLD runs were slice-time corrected using 3dTshift from AFNI 20160207 (Cox and Hyde, 1997). The BOLD time-series (including slice-timing correction when applied) were resampled onto their original, native space by applying the transforms to correct for head-motion. These resampled BOLD time-series will be referred to as *preprocessed BOLD in original space*, or just *preprocessed BOLD*. The BOLD time-series were resampled into standard space, generating a *preprocessed BOLD run in MNI152NLin2009cAsym space*. First, a reference volume and its skull-stripped version were generated using a custom methodology of *fMRIPrep*. Several confounding time-series were calculated based on the *preprocessed BOLD*: framewise displacement (FD), DVARS and three region-wise global signals. FD and DVARS are calculated for each functional run, both using their implementations in *Nipype* (following the definitions by (Power et al., 2014). The three global signals are extracted within the CSF, the WM, and the whole-brain masks. Additionally, a set of physiological regressors were extracted to allow for component-based noise correction (*CompCor*, (Behzadi et al., 2007). Principal components are estimated after high-pass filtering the *preprocessed BOLD* time-series (using a discrete cosine filter with 128s cut-off) for the two *CompCor* variants: temporal (tCompCor) and anatomical (aCompCor). tCompCor components are then calculated from the top 5% variable voxels within a mask covering the subcortical regions. This subcortical mask is obtained by heavily eroding the brain mask, which ensures it does not include cortical GM regions. For aCompCor, components are calculated within the intersection of the aforementioned mask and the union of CSF and WM masks calculated in T1w space, after their projection to the native space of each functional run (using the inverse BOLD-to-T1w transformation). Components are also calculated separately within the WM and CSF masks. For each CompCor decomposition, the *k* components with the largest singular values are retained, such that the retained components’ time series are sufficient to explain 50 percent of variance across the nuisance mask (CSF, WM, combined, or temporal). The remaining components are dropped from consideration. The head-motion estimates calculated in the correction step were also placed within the corresponding confounds file. The confound time series derived from head motion estimates and global signals were expanded with the inclusion of temporal derivatives and quadratic terms for each (Satterthwaite et al., 2013). Frames that exceeded a threshold of 0.5 mm FD or 1.5 standardized DVARS were annotated as motion outliers. All resamplings can be performed with *a single interpolation step* by composing all the pertinent transformations (i.e. head-motion transform matrices, susceptibility distortion correction when available, and co-registrations to anatomical and output spaces). Gridded (volumetric) resamplings were performed using antsApplyTransforms (ANTs), configured with Lanczos interpolation to minimize the smoothing effects of other kernels (Lanczos, 1964). Non-gridded (surface) resamplings were performed using mri\_vol2surf (FreeSurfer).

**Region of interest selection**

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

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

**Multivariate pattern analysis**

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

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

**Whole-brain searchlight.**

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

**Statistical analyses.**

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

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

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