October 26, 2021

Dear Dr. Martin,

We are thrilled to be pre-accepted at Current Biology. In response to your comments, we have edited the main body of the manuscript down to 5,543 words, completed the STAR checklist, and included the additional behavioral results asked for by Reviewer #3 in a new supplemental figure. We hope that these changes make our manuscript satisfactory for publication. Attached is the original response to reviewers for reference.

Sincerely,

Augustin C. Hennings, Mason M. McClay, Michael R. Drew, Jarrod A. Lewis-Peacock, and Joseph E. Dunsmoor

September 28, 2021

Dear Dr. Martin,

We would like to thank the Reviewers for their detailed comments and supportive statements regarding our manuscript. We have incorporated their feedback in our revised manuscript, and have responded to each of their comments below. The changes we have made in light of the Reviewers' comments have improved our submission. We are also happy to address any additional comments, or to clarify our responses if we did not address a concern to the Reviewer's satisfaction.

Sincerely,

Augustin C. Hennings, Mason M. McClay, Michael R. Drew, Jarrod A. Lewis-Peacock, and Joseph E. Dunsmoor

---

Reviewer #1: "This is a very clearly written manuscript. The data were presented clearly and the figures look great. However, there is very little novelty in the findings, aside from the distinction between anterior vs. poster hipp. The senior author is very well-aware of the substantial literature already in existence showing the distinction between the dACC and the vmPFC- only one of those studies was cited. None of the prior reviews or the prior literature that have already established the relationship between IL/PL and vmPFC/dACC has been cited. In summary, the manuscript is wonderful and great- but most of it is not novel. The literature is not properly reviewed or represented."

Response: We regret that we did not provide a better and more vigorous background for the relevant literature that frames our study, and we acknowledge that we did not cite enough of the seminal work on the distinction between the dACC and vmPFC. As the Reviewer notes, there is a body of work that shows the dissociation between the dorsal and ventral medial PFC in the context of fear-conditioning and extinction (e.g., work by Milad, Phelps, LaBar, Kalisch, Liberzon, and others). We regret failing to provide a clear explanation that our results are an extension of this important line of work. Indeed, we focused our analysis predominantly on the dACC and vmPFC because of this prior research including rodent data showing IL/PL distinctions and human data showing vmPFC/dACC distinctions. One of our major findings, and where we feel the innovation in our study resides, is showing the use of a fMRI pattern similarity approach to reveal that conditioned stimuli (CS) encoded in two separate temporal contexts (fear-conditioning and extinction) are separately reinstated (in dACC and vmPFC, respectively) within the same subject as a function of when that CS was encoded. A key feature of our protocol was the use of a hybrid design, borrowing from both associate learning and episodic memory paradigms, that used trial-unique stimuli (i.e., each CS appeared only once during either fear-conditioning or extinction). This design allowed us to evaluate fMRI data grouped not only by CS type (i.e., CS+, CS-), but also by temporal context (fear-conditioning, extinction) from when the CS was encoded. In our revision, we highlight the innovation of our work, and we have worked hard to remedy our prior shortcomings by properly grounding our findings in the substantial literature that came before.

Reviewer #1: "Along the same lines, the authors only cite studies that are conveniently supporting their findings but do not other studies that do not."

Response: We regret that there appears to be selective citations for a particular literature or set of previous findings. This is not our intention. Without knowing specifically which findings and citations the reviewer is referring to, we may assume it is in reference to studies showing dissociations between the dorsal and ventral mPFC in fear-conditioning versus extinction. Much of the human fMRI literature in this area suggests a role for the vmPFC in successful extinction recall, and abnormal univariate activity in the vmPFC in PTSD versus comparison groups.

However, a recent meta-analysis by Fullana et al. (Neuroscience & Biobehavioral Review, 2018) found that these results were only strongly observed with the use of a CS+extinguished versus CS+unextinguished design. As a whole, the vmPFC does not survive a meta-analysis of human fear extinction studies, which mostly incorporate a single repeated CS+ (and CS-). This is an important point that we have highlighted more in this revision. However, we do not interpret this as contradictory evidence to our findings. Rather, we argue that this discrepancy shows that our multivariate analytical approach, combined with our hybrid experimental approach, is a successful way to translate data from rodents to humans on the role of the vmPFC in extinction learning. Indeed, in our own data, a univariate analysis that averages activity across voxels did not reveal the vmPFC to be strongly active during extinction or during the memory test for extinction-specific CS+ items, consistent with the meta-analytic findings from Fullana et al.

If there are any other citations that we should include that would be at odds with our interpretation, we are happy to include and discuss them.

Reviewer #1: "The imbalance of the circuit in PTSD has also been published and such studies again not referenced."

Response: We regret not providing more context on the background literature on fMRI studies of PTSD. The reviewer rightly points out that there are several publications showing dysregulated or abnormal activity in the vmPFC and other brain regions in PTSD versus comparison groups during extinction learning and retrieval. We have added to our revision to highlight this important literature.

What we feel is an important and novel finding, specific to our results, is that individuals with PTSD symptoms show a similar long-term memory organization for CS+'s associated with the temporal context of fear-conditioning—that is, the fMRI pattern similarity between encoding and retrieval for CS+ items from conditioning was, as in healthy adults, expressed in the dACC, posterior hippocampus, and insula. However, despite what appeared to be successful within-session extinction learning in the PTSD symptom group, there was not a similar organization of CS+ items in the vmPFC or anterior hippocampus as found in healthy adults. Rather, the PTSD group showed reactivation of CS+ items encoded during extinction in the dACC, just as they did for CS+ items from conditioning.

Altogether, we feel this is an important new methodology to assess the long-term integrity of extinction memory in clinical groups versus healthy comparison groups. In particular, our approach is effective for investigating the spatial organization of memory specific to fear-conditioning and fear-extinction using fMRI. As others have long noted, it is adaptive to maintain memories of fear following extinction learning, as the fear associations may once again become relevant. Our results suggest that healthy individuals show this type of organization, segregating experiences between dissociable regions of the mPFC based on the temporal context in which they were acquired. Individuals with PTSD symptoms, on the other hand, show a disorganized pattern of reinstatement, such that stimuli encountered during either conditioning or extinction are similarly reinstated in the dACC. This could suggest that despite what appears to be intact extinction learning (assessed via physiological, shock expectancy, and other behavioral measures) may be an artefact of utilizing a (likely) maladaptive neural circuit that bypasses laying down a long-term memory trace of extinction in the vmPFC and anterior hippocampus.

Reviewer #2: Hennings and colleagues ask whether experiences of fear and extinction are stored as distinct memory traces, and specifically whether they are stored in distinct regions of the PFC. They further test whether this distinction differs between participants with PTSD and healthy adults and probe potential links to MTL regions. They present a clear, elegant study, in a well-written manuscript that has a strong theoretical foundation. I have only two major comments

Response: We thank the Reviewer for the positive feedback.

Reviewer #2: 1. Selective reinstatement was calculated by correlating the pattern of each trial with the equivalent pattern at retrieval, then comparing conditions by averaging the correlations. This analysis could tap into effects that are more global, and do not necessarily reflect reinstatement. For example, let us say (simplistically and unrealistically) there is some specific pattern of 'fear' in dACC. And that this 'fear' pattern is active at encoding and retrieval. There would be stronger correlations for all CS+ compared to CS- trials in dACC, but this would just be driven by this general fear pattern. I believe a stronger analysis, that would take full advantage of the experimental design, would be to compare correlations to random pairings (within category). One would calculate the difference in the average correlation of conditions, as the authors do, but assess significance by comparing it to the average difference when randomly pairing encoding trials with retrieval trials of other exemplars of the same category. This can be done in all analyses, including the mixed models, and would provide stronger evidence of selective reinstatement.

Response: We appreciate the Reviewer's in-depth thoughts on our analyses, and the opportunity to strengthen our manuscript.

To be upfront, after much discussion amongst ourselves we are not 100% certain if what we provide below as our response speaks directly to the Reviewer's comment. We took two approaches to address this comment but are open to providing more information if we somehow missed the heart of the matter.

The clearest way we read this comment was that there is a question whether there is a generalized effect of encoding-retrieval pattern similarity across the CS+ category during fear conditioning and during fear extinction. Based on our previous work using a trial-unique category conditioning paradigm (Dunsmoor et al., 2014, *Cerebral Cortex*; Morey et al., 2020, *Neuropsychopharmacology*), we fully expected for there to be a strong degree of neural similarity for all CS+ items encoded within the same temporal context. That is, based on our prior work that focused exclusively on one-day fear conditioning, the fMRI pattern similarity of CS+ items from the same category were more similar than for CS- items in regions associated with fear-learning. The key question we sought to address with the current study was whether these generalized patterns from encoding are reinstated at retrieval. This question is one of context specificity, not item specificity. However, most of the analyses we presented involved item-to-item pairings between encoding and retrieval, and we understand how this may have inadvertently focused readers' attention on item-specific rather than context-specific effects.

We now include a new analysis in the **Supplementary Results** (and replicated below). Here, we take the similarity between each item's encoding pattern to the retrieval patterns of all other items of the same CS type from the same encoding phase. For example, the encoding pattern of a single CS+ item from fear conditioning is correlated with the retrieval patterns of different CS+ items that, crucially, were encoded within the same temporal context. These values were then submitted to the same analysis as before, taking the difference in CS+ minus CS-similarity by encoding phase for each ROI. As expected, the results are qualitatively similar to the item-to-item approach we emphasized in our manuscript.

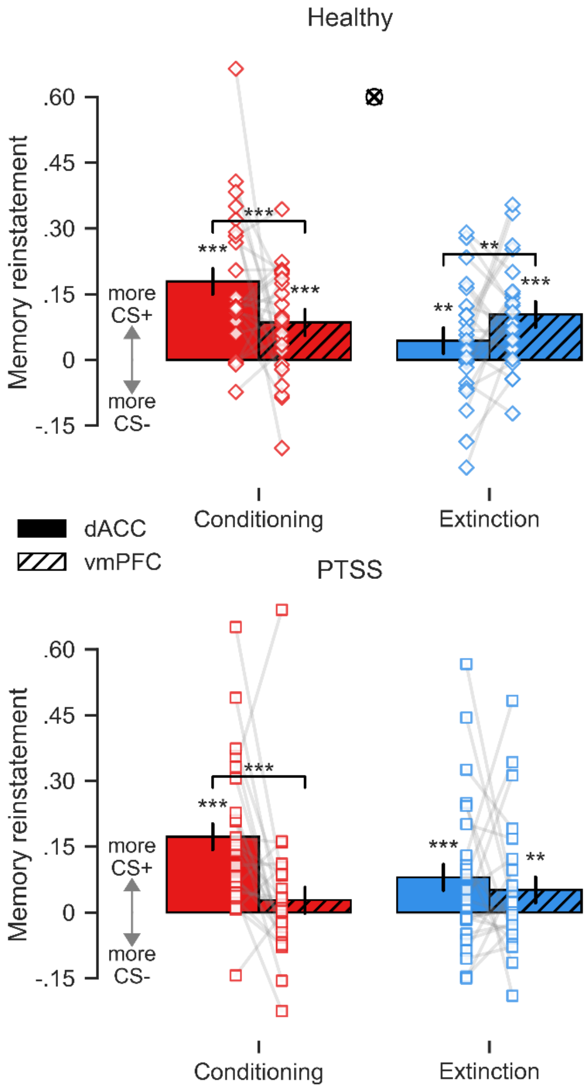
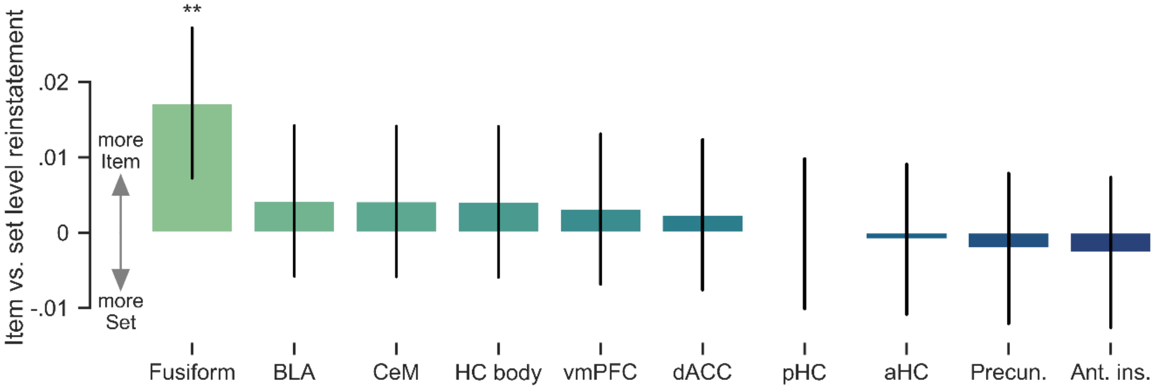


Figure S2. Dissociable reinstatement of emotional memories at the set level. All error bars correspond to the 95% confidence interval of the CS+ – CS- difference. ***P < 0.001, **P < 0.01, *P < 0.05, FDR corrected. *Top.* As in our item-to-item analysis healthy adults exhibited a significant double dissociation of set level emotional reinstatement in the mPFC, such that reinstatement for conditioning was higher in the dACC, and extinction reinstatement was higher in the vmPFC. *Bottom*. In adults with PTSS, there was no difference in extinction reinstatement between the two mPFC ROIs, suggesting that extinction memory organization is dysregulated in this group.

The crucial aspect of this finding that needs more emphasis is that the fMRI pattern similarity results in particular regions (i.e., dACC and vmPFC) are a function of the temporal context in which the items were encoded. That is, healthy adults exhibit a double dissociation in emotional memory reinstatement between the vmPFC and dACC, while individuals with PTSS do not show this double dissociation. This suggests an interesting generalization of item similarity across distinct items that were encoded *within the same temporal context* which is lacking for the PTSS group.

This new analysis can be considered at the "set" level, as opposed to the item level, as described in one of the first studies of multi-voxel pattern similarity of encoding-to-retrieval similarity (Ritchey et al., 2013, *Cerebral Cortex*). Interestingly, when looking at the "set" versus "item" level approach we broadly replicate those by Ritchey et al. (2013) showing that early visual regions are more similar at the item-level, while prefrontal regions show more generalized tuning for the "set" level, suggesting a level of abstraction. We didn't include all of these analyses in the paper (but show below), as they seem tangential to the specific goals of our paper, though they are interesting and we would be open to including them if the Reviewer feels strongly about this.



Review Figure 1. Comparison of item vs. set level reinstatement in all ROIs. An exploratory analysis comparing item vs. set level reinstatement for each item shows that item level reinstatement is higher only in the fusiform. There were no interactions with group, CS type, ROI or encoding context. These results match previously published research on cortical reinstatement. Error bars correspond to the 95% confidence interval of the item – set level difference. ** P < .01 FDR corrected.

Importantly, for the goals of our primary analysis, we are measuring reinstatement during a recognition memory test, and as such the task demands are orthogonal to those of Pavlovian conditioning where subjects are placed in a state of either anticipating or not anticipating the shock during the retrieval test (e.g., a traditional extinction recall test). For this reason, we purposefully adapted an episodic memory approach, presenting all items from all encoding trials from different CS types intermixed across trials. Thus, it would be unlikely that our results stem from a global fear pattern that is present during retrieval over many different serial presentations of CS+/- items from all encoding contexts. Rather, each time a CS+ is presented, a specific pattern corresponding to the associative memory of fear (or extinction) is reinstated based on when the CS+ item was encoded.

Another possible way we interpreted the Reviewer's comment was whether there is a general pattern of activity for *all CS+ trials*, regardless of when (conditioning, extinction) the particular item was encoded, and whether this general similarity is driving our observed results. The take-home point of our paper is to suggest that the dACC and vmPFC (among other regions reported) separately and selectively organize CS+'s as a function of the temporal context at encoding (conditioning or extinction). But is this truly selective to the temporal context, or would we see the same pattern for all CS+'s? A "general fear pattern" in the dACC for example, as the reviewer noted, that shows correlations amongst all CS+ trials. For example, would a particular CS+ exemplar (say a dog encoded during fear conditioning) resemble the pattern of another CS+ exemplar (say a kangaroo encoded during extinction), just as strongly as the item-to-item correlation we report? And, critically, would this analysis reveal the same double dissociation we report in healthy adults?

To test this alternative hypothesis, we analyzed the mean cross-phase encoding retrieval similarity. For example, we took a single trial from conditioning during encoding and measured its similarity to the retrieval of all extinction trials. We calculated this similarity for all encoding trials during conditioning and extinction using the corresponding retrieval trials (i.e., extinction and conditioning). This analysis was run within category, and we then calculated the CS+ minus CS- difference in our two mPFC regions of interest.

In accordance with the reviewer's comment, this analysis produced a loss of selectivity for the dACC and vmPFC in healthy adults. This result suggests two things; firstly, there are shared neural responses between CS+ items encoded during conditioning and extinction in the mPFC. Secondly, this analysis supports our main hypothesis that the double dissociation we observed in the item-to-item analysis arises as a function of the temporal context in which each item was encoded, not simply CS type.

In individuals with PTSS, we also observed significant CS+ > CS- cross-phase encoding-retrieval similarity, again indicating shared neural responses for these items. However, unlike the healthy adults, this group displayed greater similarity in the dACC compared to the vmPFC for both conditioning and extinction. These results further support that in individuals with PTSS, the vmPFC is dysregulated during the encoding and retrieval of fear and extinction. We have added this analysis to our **Supplementary Results** (and replicate below)**.**

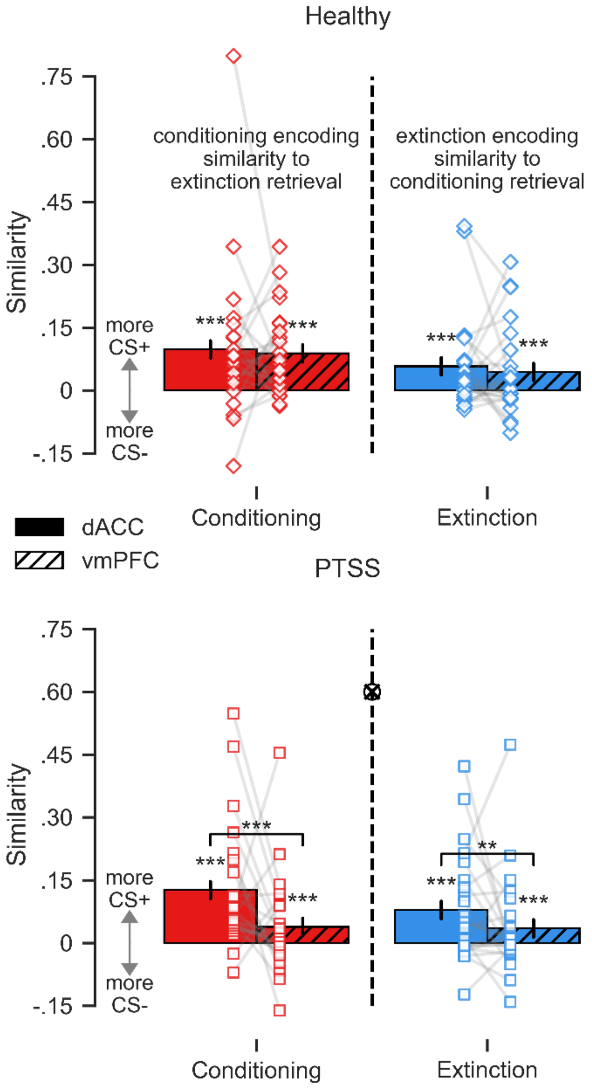


Figure S3. Cross phase encoding-retrieval similarity. All error bars correspond to the 95% confidence interval of the CS+ – CS- difference. ***P < 0.001, **P < 0.01, *P < 0.05, FDR corrected. *Top.* In healthy adults cross phase similarity does not show a double dissocation between the vmPFC and dACC, suggesting that temporal context of encoding was a key factor in our primary results. *Bottom*. In adults with PTSS, there was significantly more cross phase similarity in the dACC for both conditioning and extinction, again suggesting general vmPFC dysfunction.

We should emphasize here that the primary purpose of our experimental design was to leverage episodic item-to-item comparisons to reveal the subtle reinstatement of associative memories. Therefore, we deliberately aligned our analytical approach with the corpus of extant episodic memory encoding-retrieval similarity studies. In this way we can also match the encoding-retrieval pattern similarity for each item as a function of whether the item was remembered or forgotten, which we include in our paper. (Interestingly, whether an item was remembered or forgotten does NOT affect the results from the dACC and vmPFC, suggesting that these regions separately organize distinct associative memory by temporal context even for items that aren't explicitly remembered. This is pretty cool).

We should also emphasize the richness of this dataset lends itself to a number of ancillary analyses of interest. For example, we also include novel foils related to the CS+ and CS- category presented during retrieval. There is a lot that we wanted to pack into this paper, but it was getting to be overcrowded. As a final note, we are making all of our data publicly available.

Reviewer #2: 2. Why was reinstatement calculated regardless of memory performance? It seems more reasonable to expect reinstatement specifically for remembered trials. If it is because there were too few trials to focus on remembered ones, this issue could still be addressed to some extent. First, in terms of analysis - by equating the trial numbers in the basic reinstatement analysis and adding a predictor of memory to the mixed models. And second, discussing the significance of reinstatement in the absence of memory.

Response: We did include an analysis relating reinstatement in our ROIs to recognition memory performance in the Supplementary Results of our original submission. However, the Reviewer has raised an important issue that we did not properly reference this analysis in the main Results. To briefly summarize these results, recognition memory did not influence reinstatement in the mPFC, however it did modulate reinstatement in the hippocampus, and to a lesser extent in the amygdala. That memory performance does not track with reinstatement in the mPFC is actually consistent with previous work which shows that encoding-retrieval similarity in midline PFC regions is not associated with memory performance (Ritchey et al., 2013, *Cerebral Cortex*). As such, the lack of a link between behavioral memory performance and encoding-retrieval similarity in the present analyses is not novel and we did not emphasize this point in our manuscript. We now properly refer the reader to this result in the body of the manuscript.

Reviewer #2: Minor comments:

1. The groups did not differ in hit rates, but did they differ in false alarms? The same hit rate given a difference in false alarms would still be indicative of a behavioral difference.

Response: We now report the high confidence false alarm rates in the **Supplementary Results**, and have reproduced our analysis here for clarity:

In addition to shock expectancy, SCR, and memory hit rate we also compared false alarm rate during the recognition memory test between healthy adults and individuals with PTSS. A mixed ANOVA of high confidence false alarm rate showed a trending effect of *group* ($F_{1, 46} = 4.01$, $P = 0.051$), such that individuals with PTSS did record marginally more high confidence false alarms compared to healthy adults. There was no main effect of CS type ($P = 0.13$), and no significant CS type * group interaction ($P = 0.68$). Even though there exists this slight bias in behavioral responding to novel lures, an mixed ANOVA of high confidence corrected recognition (hit – false alarm rate) reveals no significant group differences in actual memory performance (main effect of *group*: $F_{1, 46} = 0.05$, $P = 0.83$, all interactions with *group* Ps ≥ 0.39). Note that this analysis differs slightly from the group comparison published in Hennings et al., 2021, as participants were not excluded in the current analyses for low memory performance. Thus, this trending difference in responding to novel lures during the recognition memory test is unlikely to influence our neural results, which focus solely on previously encoded probe items.

2. In figure 2 it would be useful to label the regions in the figure itself, instead of providing them only in the legend.

Response: We agree and have reorganized the figure to move the labels closer to the actual data.

3. If I understand correctly, all CS+ trials (including those followed by a shock) were treated as one condition. I assume there is not enough power for this, but if there is - it would be interesting to test for differences between trials followed by the US vs. those that are not. For example, the vmPFC may be selectively responding only to trials without a shock, regardless of whether they are CS+ or CS-, which would lead to a slightly different interpretation.

Response: We again thank the reviewer for a thoughtful analysis suggestion. We had run this analysis but left it on the cutting room floor in the initial submission. But we agree that it is an interesting and relevant analysis. There is sufficient power, as there were a 12 CS+ with and 12 CS+ without shock at fear conditioning. Importantly, we do not see any effect of US reinforcement on reinstatement in the mPFC, amygdala, or hippocampus. These results are now reported in the **Supplementary Results** section. This provides solid evidence that the central fMRI pattern similarity effects are selective to the temporal context (i.e., the CS+'s were encoded during the conditioning phase), and not specific to whether the CS+ item itself was paired with shock. This in some ways dovetails with our prior findings (Dunsmoor et al., 2018, Nat. Human. Behav.) using the category conditioning paradigm that selective recognition memory effects for CS+ versus CS- exemplars are not affected by whether the CS+ item was paired with shock or not. In effect, the associative value generalizes within the category. But the temporal context still plays an important role in segmenting long-term memory representations as a function of whether the item is associated with conditioning or extinction.

4. I was surprised to see the division of the amygdala to subfields with a standard 3mm resolution - if the authors have references demonstrating the feasibility of dissociating the two those would be helpful. Although not critical as this is a relatively minor part of the paper.

Response: Reviewer #3 raised a similar concern. Importantly, our analysis pipeline relies on robust anatomical segmentation with the 1mm^3 resolution T1-weighted images as inputs. These high-resolution anatomical masks are then interpolated into functional resolution in such a way that they do not overlap with each other (multi-label nearest-neighbor interpolation). We have clarified our description of this process in the methods.

Reviewer #3: This paper is of high interest to a broad audience of neurobiologist, psychologists and clinical scientist, as it aims to delineate key pathways that encode danger (threat) vs. safety (extinction) memories in humans. By employing an elegant experimental design, the authors show that retrieval of threat memories involve the dorsal ACC and the posterior hippocampal subfield, as well as connection between the dACC and the central nucleus of the amygdala. In contrast, extinction memories involve activation of the ventromedial PFC and connection to the anterior hippocampus. Interestingly, such clear dissociation between threat and extinction memories is not found in individuals that exhibit symptoms of post-traumatic stress. The paper is well written, concise and the analysis are fine-grained and state-of-the-art.

Response: We greatly appreciate the positive feedback.

I have two major points:

The authors could conduct group comparisons of the double dissociation in memory reinstatement within the mPFC and hippocampal subfields in order to make statistical valid statements of group differences. Additionally, the authors need to state clearly how the current manuscript differs from the earlier publication by Hennings et al 2020.

Here are my comments in detail:

Methods:

1) The authors state in the behavioral results (page 8, line 162) that only trials with high confident hits are used for analysis. This is a different analyses approach when compared to methodology in earlier publications using the same experimental design (Dunsmoor et al. 2015 Nature). Furthermore, the analysis of multi-voxel fMRI activity patterns to identify similarity between encoding and retrieval now includes all trials "irrespective or memory performance" (page 11 line 192). The authors need to clarify this duality in trial analysis

and make a clear statement why they choose to include different trails in different analyses.

Response: The Reviewer notes that the behavioral memory analysis focused on high-confidence hits, whereas one of the senior author's prior publications collapsed high and low confidence memory. First it is important to note that the memory results in this manuscript were presented primarily to demonstrate that a similar number of items were remembered between the healthy and PTSD groups, and it was not a goal of this paper to delve too deeply into the behavioral findings. However, we additionally note that the behavioral results were of strong empirical interest, and that we have analyzed memory performance and report our findings in a recent manuscript (Hennings, Lewis-Peacock, Dunsmoor, 2021, Learning & Memory). We focus in that report on memory from the pre-conditioning phase, showing that we replicate retroactive memory enhancements for CS+ items as reported in Dunsmoor et al., (2015). Additionally, we report in that manuscript that these retroactive memory effects are related to source memory misattributions of the temporal context of encoding, such that subjects who mistakenly source CS+ items encoded before fear conditioning as having been seen during fear conditioning show stronger selective retroactive memory.

Secondly, and perhaps more directly to the Reviewer's comment, the behavioral results are not affected by whether we focus on high-confidence versus collapsed (high + low confidence) memory, in regards to our two groups being matched in memory performance. We have added **Table S2**, which contains the full breakdown of memory performance at all confidence intervals for each phase for both groups.

Finally, the question of focusing the MVPA analysis irrespective of memory performance was raised by Reviewer #2 as well: We did include an analysis relating reinstatement in our ROIs to recognition memory performance in the **Supplementary Results** of our original submission. However the Reviewer has raised an important issue that we did not properly reference this analysis in the main Results. We have remedied this. To briefly summarize these results, recognition memory did not influence reinstatement in the mPFC, however it did modulate reinstatement in the hippocampus, and to a lesser extent in the amygdala. That memory performance does not track with reinstatement in the mPFC is actually consistent with previous work which shows that encoding-similarity in midline PFC regions is not associated with memory performance (Ritchey et al., 2013, *Cerebral Cortex*). As such, the lack of a link between behavioral memory performance and neural reinstatement in the present analyses is not novel.

2) I understood that contrast images (CS+ - CS-) have been used for the searchlight analysis (page 23, line 681). However, the authors wrote that "selective CS+ reinstatement" was examined (page 13, line 245). Please comment.

Response: In the context of this result, our use of the term “selective reinstatement” refers to the significant CS+ > CS- difference we observed. However, we acknowledge that this language is ambiguous, and we have revised this section and others to include more precise descriptions of our effects of interest.

3) Could the authors illustrate that the resolution of 3 cubic mm (plus smoothing see the next point) is sufficient to investigate separate nuclei in the amygdala?

Response: Reviewer #2 raised a similar concern. Importantly, our analysis pipeline relies on robust anatomical segmentation with the 1mm^3 resolution T1-weighted images as inputs. These high-resolution anatomical masks are then interpolated into functional resolution in such a way that they do not overlap with each other (multi-label nearest-neighbor interpolation). We have clarified our description of this process in the methods.

4) I am no expert for the fMRIprep pipeline, but is it possible to state the smoothing of the functional MRimages?

Response: Functional smoothing was purposefully omitted from all of the analyses presented in manuscript. Since our regions of interest are either *a priori* ROIs from meta-analyses, or anatomical segmentations, functional smoothing would introduce noise from neighboring voxels outside of our ROIs, at the cost of analytical specificity. This would be especially problematic for interpreting our hippocampal and amygdalar results. We note that the primary reason to smooth images is to increase the signal-to-noise ratio when conducting univariate GLMs. However, as our principle analyses are multivariate in nature, functional smoothing is not necessary. We have retained this information in the manuscript, page 32: "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."

5) The prediction of reinstatement in the mPFC ROIs by MTL activity is an intriguing idea. However, the regression model does not predict the activation in the vmPFC or the dACC per se, but their difference (which is not directly clear from the figure 4). While I understand this differential contrast is chosen, the authors might consider to include a connectivity analysis that are established for fMRI analyses (e.g., psycho-physiological interaction in SPM).

Response: We thank the reviewer for their interest in this analysis. To begin, we have edited Figure 4 to better clarify that the purpose of this analysis is to predict the difference in reinstatement between the dACC and vmPFC. This analysis assesses how MTL regions bias reinstatement towards one mPFC region and away from the other, which would not be captured by traditional functional connectivity analyses. PPI assess the change in co-activation between two regions as a function of task, but our outcome variable of interest in this case is multivariate reinstatement, not average univariate activity. As such, we feel that adding another type of complex analysis (e.g., functional connectivity) would serve to overcrowd the manuscript while not allowing us to answer the question of what biases reinstatement in the mPFC.

Results:

6) The authors report no differences between groups in behavioral and psychophysiological results during the renewal test, but report differences between groups in Hennings et al. 2020 Neuropsychologica. Please comment.

Response: There is likely some confusion, which we hopefully cleared up in the revision, between the fear renewal test and the recognition memory test. These are two separate phases on Day 2 composed of different sets of stimuli. In the current manuscript, we focus exclusively on Day 1 (encoding) and the recognition memory test from Day 2. Fear renewal was presented in Hennings et al., (2020, Neuropsychologia). The fear renewal test preceded the recognition memory test on Day 2, but this was not the focus of the current manuscript and thus the data are not included here. The purpose of presenting behavioral results from Day 1 and recognition memory results from Day 2 in this manuscript was to show that the healthy and PTSS groups were matched on initial learning (SCRs and shock expectancy during conditioning and extinction learning) and matched on the average number of items remembered from each category during each phase.

7) With respect to the different results in healthy volunteers and individuals with PTSS in encoding-retrieval similarity between threat and extinction memories (page 12 and following): Is there a difference between groups? While the methods description states

that group is a factor in the multiple regression model (page 33, line 712), the authors only report that there is no significant CStype * encoding context *ROI interaction in the PTSS group (Page 13, line 239). Is it possible to do a group comparison to make the group difference statistically valid? Similarly, is it possible to make group comparisons for the hippocampal subfield analysis? Thereby the authors could statistically strengthen their statements in the discussion (e.g., page 21, line 410 on "different […] neural mechanisms").

Response: We thank the reviewer for the opportunity to include more focused statistical comparisons. For both the mPFC and hippocampus, we now directly test the pattern of observed reinstatement for both fear conditioning and extinction memories in the main body of the manuscript. As expected, there are significant group differences in both the mPFC and hippocampus for the pattern of extinction memory reinstatement, but no group differences for fear memory reinstatement.

8) How would encoding-retrieval pattern look like for the anterior insula? I see that the authors have their predefined ROIs and that reviewer should not let the author to search for several additional ROIs, but I think it might be informative to the broader readership to clarify the role of the AI. This role might be based on the cited meta-analysis by Fullana et al., as well as the searchlight analysis.

Response: We thank the reviewer for the opportunity to explore these interesting results in more detail. While attempting to avoid over-filling the paper with analyses, we have added an analysis of reinstatement in both the anterior insula, as the reviewer suggested, and the precuneus. The anterior insula, like the dACC, is another region that is consistently identified in fMRI analyses of fear conditioning and extinction, and it may contribute to the experienced emotional content of an event. The precuneus has been consistently implicated in episodic retrieval, and as such is another interesting location to probe the differential reinstatement of these emotional associations. We found significant CS+ > CS- reinstatement in both regions, and now include these results and discuss their implication in the revised manuscript.

Minor:

-Page 4, line 74: There are a few publications on vmPFC involvement in extinction retrieval, e.g., Milad et al. Proc Natl Acad Sci U S A. 2005, Biol Psychiatry. 2007, Kalisch et al. J Neurosci. 2006

Response: We completely agree and regret not including more of the seminal work in our manuscript. This has been fixed.

-Page 12, line 211: "there was a selective reinstatement". How was the reinstatement selective, if the vmPFC was active during retrieval of both, threat and extinction memories?

Response: We have specified our discussion of this result to clarify that we were referring to greater CS+ reinstatement as compared to CS- reinstatement. In general, we have reframed our discussion of this CS+ specific effect, so as not to confuse it with the various ROI specific results. For example, in healthy adults, extinction reinstatement was selective to the vmPFC relative to the dACC.

-Page 20, line 402: The authors might reconsider the term "abnormal"

Response: We agree and thank the reviewer for catching this. We have removed the word “abnormal” from the manuscript.