Reviews

Reviewed by Tom Beckers, 02 Jun 2023 20:31

The authors have done a very careful and thorough job in addressing the reviewer's comments (mine but also those of the other 2 reviewers, as far as I can judge) and revising the manuscript. I do not see many remaining issues. One thing that still stands out, though, is the ambivalence regarding what the effect of instructions should be expected to be for the extinction of aversive versus appetitive learning. This is evident already in the abstract: the authors write initially that 'In a therapeutic context, this asymmetry [i.e., aversive learning being slower to extinguish than appetitive learning] that has been discussed as indicative of a 'better safe than sorry strategy' could potentially be overcome by making patients aware of the change in contingency' [implying that instructions should make extinction of aversive cues more like extinction of appetitive cues]. Yet a few lines down, they state that 'We expect [that the] effect [of instruction] is more pronounced for appetitive stimuli [than for aversive stimuli], which implies that instructions would further enhance the asymmetry between extinction of appetitive and aversive cues in the context of pain. It is also the latter prediction that they subscribe to in their response to point 3 of my initial review. Here they refer to the same 'better safe than sorry strategy' that is supposed to yield the initial asymmetry between extinction of appetitive versus aversive cues as the reason to expect a lesser effect of instruction, whereas the opening sentences of the abstract suggest that instructions are expected to counter exactly the effects of that strategy.

This ambivalence regarding the anticipated effect of instructions shines through elsewhere also. E.g., on p. 4, line 6-9, it is again suggested that instructions may be a promising method to prevent the incomplete extinction of aversive associations (implying that it would bring aversive extinciton more in line with appetitive extinction]. This conceptual confusion deserves wrinkling out, I think.

Other than that one issue, however, I think that the authors have done an excellent job and have dealt with all the comments in a satisfactory way. I would be happy to see this registered report advance to stage 2.

Tom Beckers

We thank the reviewer for the positive feedback and are very grateful for the thorough evaluation of our manuscript which indeed contained an inconsistency in the predictions of the effect of instructions that had escaped our attention. In the following, we would like to clarify our hypotheses and offer a revised wording.

As explained in the manuscript, the comparison between extinction following aversive and appetitive learning is motivated by previous observations showing that extinction following aversive learning is more resistant to extinction (indexed by slower or incomplete extinction, i.e., a still stronger conditioned response at the end of the extinction phase). We refer to this observation as indicative of a "better safe than sorry" strategy and expect to find a similar pattern in our uninstructed condition (see hypothesis H2). This should be reflected in a shallower slope and a higher CR at the end of the extinction phase.

Regarding the effect of instructions, we next hypothesise that instructions facilitate extinction and should therefore lead to a steeper slope and a weaker CR at the end of extinction compared to the uninstructed condition (i.e., interaction: *instruction group* (instructed, uninstructed) × *time* interaction, hypothesis H3). Importantly, this effect should not only occur *across* aversive and appetitive learning, but should also be significant for each of them individually.

However, as in the uninstructed condition, the effect of instructions is expected to differ between the aversive and appetitive conditions in the sense that instructions lead to more extinction in the appetitive condition (interaction: *CS type* × *instruction group* × *time* interaction; hypothesis H4).

Together, hypotheses H3 and H4 therefore formalise the assumptions that instructions can facilitate extinction (H3), albeit to different extents in the aversive and the appetitive conditions (H4).

The notion of enhanced extinction also in the aversive condition had led us to write that instructions would overcome the asymmetry between the aversive and appetitive condition. However, this is not correct. Instead, the enhancement effect of instructions *increases* the asymmetry due to the differential effect on the two types of learning, as also pointed out by the reviewer.

To reflect the point, we revised the relevant sentence in the abstract as follows:

"In the context of pain, extinction learning has been shown to be slower or incomplete for aversive compared to appetitive cues (i.e., cues signaling pain exacerbation and pain relief, respectively), potentially due to their higher biological relevance. In a therapeutic context, this reluctant extinction that has been discussed as indicative of a 'better safe than sorry strategy' could potentially be reduced by making patients aware of the change in contingency."

We would like to thank the reviewer again for pointing out this inconsistency and hope that this clarification and revision of the manuscript will resolve his point.

Reviewed by Karita Ojala, 19 May 2023 15:15

After going through the revised manuscript and author replies in detail, I am happy with how the authors have responded to all of the reviewer comments and revised the manuscript accordingly. I commend the authors for their thorough work.

The only remaining point I have pertains to the definitions of the regions-of-interest (ROIs) for the fMRI analyses, which I unfortunately overlooked in my first review - my apologies.

The manuscript states (p. 30, line 31-33 on tracked changes version): "SBFC analyses will use the left and right dlPFC and vmPFC, amygdala, and striatum as seeds, with masks derived from the FSL Harvard-Oxford Atlas...".

However, dlPFC, vmPFC, or striatum do not exist as regions in the Harvard-Oxford Atlas (amygdala does). I would ask the authors to precisely define the regions from the Harvard-Oxford Atlas (or another atlas) that will be used to construct the specified ROIs, or to offer an alternative definition of the ROIs and a description of how the masks will be built.

As a side note of this point, striatum is mentioned in the manuscript for the first time as "ventral striatum" but all other mentions are to "striatum" only, so I am not certain if this means both ventral and dorsal striatum together, or is only used as a short form for ventral striatum - please clarify. Based on the extensive evidence that ventral and dorsal striatum have different functions, shown also in many fMRI conditioning/learning studies, it would make sense to differentiate them and (re)consider whether to include e.g. only ventral striatum, or both.

We thank the reviewer for the positive feedback and the opportunity to clarify the definition of these three regions of interest. As the dlPFC and vmPFC are indeed not defined in the Harvard-Oxford Atlas (Desikan et al., 2006), these two regions will be defined based on the meta-analysis tool in Neurosynth (Yarkoni et al., 2011) that synthesizes previously published imaging data. We will extract masks for the named ROIs from a whole brain functional parcellation.

Regarding the striatum, we agree with the reviewer that due to their different functional roles in learning, differentiating between the dorsal and ventral striatum is of interest in our planned analyses (see page 7 for highlighted changes). We now specify on pages 30 to 31 that masks will be generated by combining the striatal subregions, i.e., caudate, putamen, and nucleus accumbens, as specified in the Harvard-

Oxford Atlas (Desikan et al., 2006). Dorsal and ventral striatum will be divided at *z* = 0 (MNI coordinates), according to Cooper et al. (2012).

"SBFC analyses will use the left and right dlPFC and vmPFC, amygdala, and striatum as seeds for first-level analyses in the CONN toolbox. Amygdala and striatum masks will be derived from the Harvard-Oxford Atlas (Desikan et al., 2006), with the striatal masks constructed by combining the caudate, putamen, and nucleus accumbens. Separate masks for the dorsal and ventral striatum will be created by a split at MNI coordinate z = 0, according to Cooper et al. (2012). Subsequent analyses will be performed for the ventral and dorsal striatum separately. Masks for the vmPFC and dlPFC will be obtained from a whole brain functional parcellation (https://identifiers.org/neurovault.collection:2099) created with Neurosynth's meta-analysis tool (Yarkoni et al., 2011)."

Reviewed by Gaëtan Mertens, 08 May 2023 09:41

I'm satisfied with the authors' replies to my comments and I look forward to seeing the results of this RR.

We would like to thank the reviewer for the constructive criticism that has clearly improved our study protocol.

References

Battaglia, S., Harrison, B. J., & Fullana, M. A. (2022). Does the human ventromedial prefrontal cortex support fear learning, fear extinction or both? A commentary on subregional contributions. *Molecular Psychiatry*, *27*(2), 784–786. https://doi.org/10.1038/s41380-021-01326-4

Becerra, L., Navratilova, E., Porreca, F., & Borsook, D. (2013). Analogous responses in the nucleus accumbens and cingulate cortex to pain onset (aversion) and offset (relief) in rats and humans. *Journal of Neurophysiology*, *110*(5), 1221–1226. https://doi.org/10.1152/jn.00284.2013

Belleau, E. L., Pedersen, W. S., Miskovich, T. A., Helmstetter, F. J., & Larson, C. L. (2018). Cortico-limbic connectivity changes following fear extinction and relationships with trait anxiety. *Social Cognitive and Affective Neuroscience*. https://doi.org/10.1093/scan/nsy073

Cooper, J. C., Dunne, S., Furey, T., & O'Doherty, J. P. (2012). Human Dorsal Striatum Encodes Prediction Errors during Observational Learning of Instrumental Actions. *Journal of Cognitive Neuroscience*, *24*(1), 106–118. https://doi.org/10.1162/jocn_a_00114

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., & Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*(3), 968–980. https://doi.org/10.1016/j.neuroimage.2006.01.021

Doll, B. B., Jacobs, W. J., Sanfey, A. G., & Frank, M. J. (2009). Instructional control of reinforcement learning: A behavioral and neurocomputational investigation. *Brain Research*, *1299*, 74–94. https://doi.org/10.1016/j.brainres.2009.07.007

Fullana, M. A., Harrison, B. J., Soriano-Mas, C., Vervliet, B., Cardoner, N., Àvila-Parcet, A., & Radua, J. (2016). Neural signatures of human fear conditioning: An updated and extended meta-analysis of fMRI studies. *Molecular Psychiatry*, *21*(4), 500–508. https://doi.org/10.1038/mp.2015.88

Klein, S., Kruse, O., Tapia León, I., Van Oudenhove, L., van 't Hof, S. R., Klucken, T., Wager, T. D., & Stark, R. (2022). Cross-paradigm integration shows a common neural basis for aversive and appetitive conditioning. *NeuroImage*, *263*, 119594. https://doi.org/10.1016/j.neuroimage.2022.119594

Leknes, S., Lee, M., Berna, C., Andersson, J., & Tracey, I. (2011). Relief as a Reward: Hedonic and Neural Responses to Safety from Pain. *PLoS ONE*, *6*(4), e17870. https://doi.org/10.1371/journal.pone.0017870

Leknes, S., & Tracey, I. (2008). A common neurobiology for pain and pleasure. *Nature Reviews Neuroscience*, *9*(4), 314–320. https://doi.org/10.1038/nrn2333

Martynova, O., Tetereva, A., Balaev, V., Portnova, G., Ushakov, V., & Ivanitsky, A. (2020). Longitudinal changes of resting-state functional connectivity of amygdala following fear learning and extinction. *International Journal of Psychophysiology*, *149*, 15–24. https://doi.org/10.1016/j.ijpsycho.2020.01.002

Milad, M. R., & Quirk, G. J. (2012). Fear Extinction as a Model for Translational Neuroscience: Ten Years of Progress. *Annual Review of Psychology*, *63*(1), 129–151. https://doi.org/10.1146/annurev.psych.121208.131631

Oldham, S., Murawski, C., Fornito, A., Youssef, G., Yücel, M., & Lorenzetti, V. (2018). The anticipation and outcome phases of reward and loss processing: A neuroimaging meta-analysis of the monetary incentive delay task. *Human Brain Mapping*, *39*(8), 3398–3418. https://doi.org/10.1002/hbm.24184

Sescousse, G., Caldú, X., Segura, B., & Dreher, J.-C. (2013). Processing of primary and secondary rewards: A quantitative meta-analysis and review of human functional neuroimaging studies. *Neuroscience & Biobehavioral Reviews*, *37*(4), 681–696. https://doi.org/10.1016/j.neubiorev.2013.02.002

Seymour, B., O'Doherty, J. P., Koltzenburg, M., Wiech, K., Frackowiak, R., Friston, K., & Dolan, R. (2005). Opponent appetitive-aversive neural processes underlie predictive learning of pain relief. *Nature Neuroscience*, *8*(9), 1234–1240. https://doi.org/10.1038/nn1527

Wendt, J., & Morriss, J. (2022). An examination of Intolerance of Uncertainty and contingency instruction on multiple indices during threat acquisition and extinction training. *International Journal of Psychophysiology*, *177*, 171–178. https://doi.org/10.1016/j.ijpsycho.2022.05.005

Yarkoni, T., Poldrack, R. A., Nichols, T. E., Van Essen, D. C., & Wager, T. D. (2011). Large-scale automated synthesis of human functional neuroimaging data. *Nature Methods*, *8*(8), 665–670. https://doi.org/10.1038/nmeth.1635