This is a stage 1 proposal for a meta-analysis on the action-effect (i.e., the finding that people experience stronger emotions for events that were the result of action as opposed to inaction). The authors propose to investigate the strength of the action effect for positive and negative emotions and for counterfactual thoughts. They also propose to test a number of potential moderators that have been identified in the literature.

Let me start by saying that literature on the action effect is not one I am familiar with. As a result, I cannot comment much on the literature review. From an outsider's perspective, I thought the literature review was clear and provided a good overview of the different theories and findings in the field. A couple of (potentially naïve) questions and comments that did come to mind while reading were:

1. On P. 6, the authors write that they focus on emotions and counterfactual thoughts and not on other action-inaction effects, which they argue are different. It was not clear to me, however, why these other effects were different. Are they not explained by the same theories? If they are explained by the same theories, then why not include them?
2. Most moderators appeared rather atheoretical. The authors write on P. 10 that current theories are imprecise and therefore difficult to test, so I assume this explains why. Nevertheless, I was wondering if some of the theories do make different predictions about some of the included moderators? In addition, I was wondering if it would be possible to directly test norm theory by conducting a meta-analytical correlation between counterfactual thoughts and positive/negative emotions, where the latter forms a proxy for regret?

I also had the following comments on the proposed methods:

1. Why not test H1a and H1b in a single model, where "type of emotion" is included as a moderator? This will give the authors more power and would have the benefit that it also allows them to test if the action-inaction effect might be stronger/weaker for negative than for positive emotions.
2. Temporal distance: I have a sense that coding this in terms of # years will result in a very skewed distribution with a number of large outliers on the right side of the distribution. This could potentially bias this analysis. How will the authors deal with this? More generally, how realistic is the assumption that the effect of temporal distance is linear? For example, do we really expect the difference between a 1-week or a 2-week interval to matter? If not, perhaps it makes sense to code this variable categorically. Could the authors comment on this?

3. The screening procedure wasn't entirely clear to me. P. 24 mentions that the lead author will screen the papers, but P. 26-27 seem to suggest that screening will be done together.
4. The effect size computation is potentially problematic. Cohen's d can be calculated in different ways for repeated measures designs (Lakens, 2013) and it's not entirely clear how the authors will calculate it in the different scenarios they identify. Based on my reading, I fear they might be collapsing different types of cohen's *d* for the repeated measures studies. The formula based on the t-test mentioned in Table 4 suggests that they will calculate d_z, which corrects the SD for the correlation between measures. How they will calculate cohen's d from descriptive information in repeated measures designs is, however, not clearly described and if I understand their code correctly, it suggests that cohen's d will be calculated there as if it were a between-subjects study, therefore not correcting the SD for the correlation between measures. I think the authors should ensure that cohen's d is always calculated in the same way. Given that both within- and between-subject studies are included, this means that they should always calculate cohen's d without correcting the SD for the correlation between measures (Lakens, 2013). This is especially important when comparing within- and between-subject studies (H3), because otherwise any difference can be trivially attributed to the fact that in repeated measures designs, cohen's d was calculated differently.
5. Building on the above, it's not clear to me how the cohen's d calculated from binary choices relates to the cohen's d calculated from continuous variables. As mentioned above, the authors should ensure that cohen's d always means the same thing across studies. For this reason, I also think it is not a good idea to use the cohen's d reported in the analyzed papers (as the authors propose on P. 32), because it will not always be clear which type of cohen's d is reported.
6. The statistical approach could use some more explanation. For example, how do three-level models correct for confounding among moderators? Relatedly, would it make sense to report the relationships between the moderators to assess confounding? (e.g., Hofmann et al., 2010).
7. I usually use RVE to deal with effect size dependence and so am not very familiar with three-level models, but if my understanding is correct, three-level models only deal with "hierarchical dependence", not with the type of dependence arising from the same sample providing multiple effect sizes (e.g., a study reporting different measures of negative emotion; Tanner-Smith et al., 2016). This latter type of dependence strikes me as more important than the hierarchical dependence. How common do the authors estimate multiple effect sizes from the same sample will be? If common, perhaps it makes sense to use RVE

(Tanner-Smith et al., 2016) instead of averaging together those effect sizes as proposed now on P. 32.

8. The authors report a power analysis, which is great, but I was wondering whether their power analysis is appropriate for the multivariate three-level models they aim to fit.
9. I didn't understand the reported posteriori power analyses. What effect sizes are these based on? How can there be high posteriori power for a non-significant effect size?
10. Given that we have data to support when one type of publication bias correction is better than another, why not report the best type given the parameters of the data and report the other corrections in supplementary material? Reporting all types of correction next to each other gives the impression that they are all equally good, which the authors themselves say is not the case.
11. The authors request feedback on when to use random forests. Given that there is no research on this, I personally think the arbitrary threshold they propose is reasonable. Alternatively, if that makes sense, they could run a power analysis to assess the power they would have with a three-level model and then use a cut-off based on the outcome of this analysis.
12. Why include 2-level model results in the moderator section?
13. Metaforest is not a very common method (yet), so it would be helpful if the authors could provide some more guidance in interpreting the output. For example, they write "the main model indicator, R-squared (R-OOB) was -0.02". What does this mean?

Minor comments:

1. The authors often seem to drop the article when speaking of the action-inaction effect. I found this a bit awkward to read.
2. Table 2: I don't really understand what the authors mean with "associated with". That is, I have difficulties relating the "description" to the "term" here.
3. P. 22: the authors refer to table 1, but it should be table 3.
4. P. 41: experimental studies → comparison studies?
5. P. 44: "We recognize that the median power of studies is 12.7%" → of which studies?
6. P. 50: "Nine studies with between-subject design had a positive mean effect" → this sounds as if there were nine studies that had a positive effect. I would rephrase throughout the results section.

References:

1. Hofmann, W., De Houwer, J., Perugini, M., Baeyens, F., & Crombez, G. (2010). Evaluative conditioning in humans: A meta-analysis. *Psychological Bulletin*, *136*(3), 390–421. https://doi.org/10.1037/a0018916
2. Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, *4*, 1–12. https://doi.org/10.3389/fpsyg.2013.00863
3. Tanner-Smith, E. E., Tipton, E., & Polanin, J. R. (2016). Handling complex meta-analytic data structures using robust variance estimates: A tutorial in R. *Journal of Developmental and Life-Course Criminology*, *2*(1), 85–112. https://doi.org/10.1007/s40865-016-0026-5