|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  | Sean Hughes  E Sean.Hughes@UGent.be  T 0032 (0)9 264 86 49  Faculty of Psychology and Educational Sciences  Henri Dunantlaan 2  B-9000 Ghent  Belgium |
|  | |  |
|  |  |  |
| date  05/04/2021 | page  1 | our reference  Response Letter | |
|  |  |  |  |

Dear Dr. Berenbaum,

On behalf of my fellow co-authors, I would like to thank you for functioning as the action editor on our recently submitted manuscript “Deepfaked Online Content is Highly Effective in Manipulating People's Attitudes and Intentions” (PNAS 2021-04216). We would also like to thank Reviewers 1-2 for their time and effort as well.

As can be expected, we were disheartened to read about the decision to not accept our paper at PNAS. Especially given the extremely positive comments of Reviewer 1 (e.g., “this is one of the best articles I've reviewed for PNAS in a while”). And his/her belief in the real world applicability of our work (e.g., “I've attended many high-level policy-maker meetings on Deepfakes and everyone's worried about them but there's very little evidence on how people respond to them. This article is the first to examine this question systematically and it does so very well in my opinion”).

This decision seems to be mainly based on Reviewer 2’s comments. While we are sincerely thankful for that feedback, we note that it was often based on incorrect arguments about our manuscript, data, and/or analyses. We briefly respond to the issues we see with those arguments at the end of this letter, and hope you will give us an opportunity to clarify these inaccuracies.

We also appreciate your various comments. In particular, we share your desire for a study of real-world political manipulation using Deepfakes. We actually approached several AI researchers about collaborating on such a project. Every one of them were unwilling to make Deepfakes of real life political figures given the social and legal ramifications of doing so. The fact that no AI researcher would create Deepfakes of real politicians only speaks to the danger of this new technology, and adds to the urgency in bringing our work to the PNAS audience.

While a focus on political Deepfakes is certainly needed, it is only one of many situations where this technology is being used. Indeed, there are already real-world cases where novel individuals (like those used in our studies) are being used for malicious purposes. For example, the FBI cyber division just released a warning that “Malicious actors almost certainly will leverage synthetic content for cyber and foreign influence operations in the next 12-18 months” and “the FBI anticipates it will be increasingly used by foreign and criminal cyber actors for spearphishing and social engineering in an evolution of cyber operational tradecraft.” (<https://www.ic3.gov/Media/News/2021/210310-2.pdf>). The synthetic content they warn of is very similar to the content we cover in our paper.

In light of the above, as well as the real and pressing danger that Deepfakes represent, we decided to examine the power of this new technology to manipulate (automatic) first impressions of others. This is the reason we used a novel individual instead of a known one. We realize that this rationale could have been better expressed in our original manuscript, and have therefore revised our paper accordingly. These revisions also address the question of “why a white male [was chosen] as the target in the Deepfakes”. The answer is quite simple: the target of our Deepfakes was the first author, who was able to take full responsibility for the potential misuse of his Deepfaked self. We could not ensure this for any other (especially well known) individual.

Finally, there seemed to be some confusion about the core aims of our work. To be clear: our main goal was to investigate if (1) Deepfakes are good enough (not necessarily indistinguishable from genuine material) to manipulate *enough* people, *enough* of the time, to be a serious cause for public concern.

Our data repeatedly showed this to be true, and we ourselves were shocked to see how strong the effects were. This is why we ran multiple pre-registered exploratory and a separate confirmation study to ensure maximum reliability of our findings. It was only after we empirically demonstrated that (1) was true that we then turned to the separate and secondary questions of (2) are Deepfakes similar to genuine content, and (3) can people detect when they are being exposed to Deepfakes.

While (2-3) are certainly interesting, the original question (1) was always our primary concern, highly important, and why we continue to believe that a PNAS audience would find it imminently useful. Reviewer 1 understood that (1) is the key finding of our work (“the first to do so”) whereas Reviewer 2 appears to have focused on the technical details of a latter, (more academic and secondary) question (2).

Taken together, we hope that after reading our responses you will consider forwarding our letter to Reviewer 2, or in light of Reviewer 1’s overwhelmingly positive evaluation (“the best article I’ve reviewed for PNAS”), request the input of a third reviewer.

We sincerely thank you for taking the time to read our response and look forward to hearing back from you soon.

Kind Regards,

Sean Hughes

(*Corresponding author*)

**Response to Reviewer 2’s Comments**

**Reviewer 2**: However, I was surprised in studies that directly compared Deepfaked and genuine content (e.g., Experiment 2, 3, 4, 5) that there was no report of the Content (positive/negative) X Type (Deepfaked/genuine) interaction. As best I can tell from the supplement, only a main effect of Video Type was reported, which seems rather irrelevant to the key point of the study. The meta-analysis seems to directly compare the content effect sizes across genuine and faked materials, but equivalent tests did not seem to be reported for the relevant individual studies. That seemed odd given how important claims of similar effects across faked and genuine content seemed to be.

i) that there was no report of the Content (positive/negative) X Type (deepfaked/genuine) interaction.

**Authors**: The Reviewer is correct that there has to be an interaction term to take into account the various effects. In all of the analysis we performed, we have included the interaction term and so it is taken into account, e.g., for each independent variable we fit a function of the form:

ID = a + b\*content + c\*type + d\*content\*type

where a, b, c, and d are fit coefficients. For the detailed analysis, see supplementary material: ./analysis/analyses\_exp\_1-6\_exploratory.html. We have not reported the individual coefficient of the interaction term because all of our hypothesis tests comparing Deepfake vs genuine effects consider it implicitly via the posterior probability of this contrast. In short, an examination of the analyses files will demonstrate that the requested interaction was always included in our model. We are more than happy to report this if the Reviewer deems it useful.

**Reviewer 2**: ii) only a main effect of Video Type was reported, which seems rather irrelevant to the key point of the study.

**Authors**: We respectfully disagree with Reviewer 2 on what the “key point of the study” was. We set out to answer the following question, can currently available Deepfakes manipulate enough people, enough of the time, to be a significant cause for public concern. It is for this precise reason that we compared Deepfakes to genuine material while marginalizing other variables (such as content type [positive vs. negative]). Given that our paper is the very first to examine this question we consciously decided to focus on Video Type to clearly illustrate a crucial point: Deepfakes are similar to genuine material regardless of the different (positive vs. negative) content contained therein.

Nevertheless, we do acknowledge in our General Discussion that future work will need to examine the potential moderators of Deepfaking effects (e.g., like the valence of the message being communicated). Yet the core aim of this very first set of studies on the topic was to marginalize (i.e., average over) across dependent variables in order to demonstrate the generalizability of our core “Deepfake vs Genuine” claims. Thus, instead of being “irrelevant”, this analysis was “the key point of the study”.

**Reviewer 2**: iii) but equivalent tests did not seem to be reported for the relevant individual studies. That seemed odd given how important claims of similar effects across faked and genuine content seemed to be.

**Authors**: Reviewer 2 is incorrect here: Bayes Factors were preregistered for each of these equivalence tests and are reported for each individual study (see relevant analyses files and sections of the Supplementary Materials). The method of testing for equivalence was changed to non-inferiority testing in the exploratory meta-analysis and preregistered for the confirmatory study, as documented in the supplementary materials.

**Reviewer 2**: In most of the studies, detection was measured by telling people that the Deepfakes were faked and then asking people whether they had recognized that. There are a number of potential difficulties with making conclusions about answers to such questions. First, in most of the studies, participants who received genuine materials were not asked whether they thought that the materials were Deepfakes. As a result, one cannot tell whether "detection" actually means detection of fakes that would not have occurred if the material was genuine. Instead, "detection" could be identifying people who had no idea the materials were fake until they were informed, but after they were informed, they reported that they "knew it all along." Such a measure would seem highly questionable as a reflection of actual detection.

i) First, in most of the studies, participants who received genuine materials were not asked whether they thought that the materials were Deepfakes.

**Authors**: Reviewer 2 is partially correct. Their concern only applies to a subset of our exploratory studies (Experiments 4-5) and importantly, not to our preregistered, high-powered confirmatory study. Reviewer 2 themselves acknowledges this in their next point (i.e., that the issue they raise here was addressed in the confirmatory study which provides a higher standard of evidence). Indeed, this factor is one of the reasons that lead us to describe Experiments 1-5 as exploratory and Experiment 6 as confirmatory. Since all our conclusions are based on the confirmatory study, the reviewer’s comments do not apply to our conclusions.

**Reviewer 2**: In Experiment 6, the detection question did not tell people that their materials had been faked, so that wording seemed better to me. Analyses showed that people who identified fake material as fake still showed effects, but there was no comparison of those effects with people who thought the faked material was genuine. It could well be that perceiving the material as fake (at least after the fact) was related to smaller effects even if those effects were still highly significant. It also remains unclear whether people would detect these stimuli as fake without being told that some stimuli were fake and others were not and then reflecting back on what they saw.

i) Analyses showed that people who identified fake material as fake still showed effects, but there was no comparison of those effects with people who thought the faked material was genuine.

**Authors**: Reviewer 2 is once again incorrect here. First, we agree with the previous point that the awareness/detection questions differed between the exploratory experiments and our final confirmatory one. As such, to directly compare the detection rates between these two would be an apples-and-oranges comparison - hence we did not do it. A second reason that we elected to not directly compare these results is that we were not well powered for such a comparison, as this question about “detection” was never the primary purpose of our work. So any test’s result would be minimally informative.

That said, Reviewer 2 has hit on several points (i.e., that this is a new, emergent, and rapidly evolving field that is highly important, urgently needed) and highlights several questions that still need to be answered. However, they cannot all be answered in this one, and first of its kind, study. We were surprised the Deepfakes were as effective as genuine material, and this point alone is sufficient to alert the PNAS audience to the dangers and powers of this new technology. Our findings empirically demonstrate what others have long assumed to be true (Deepfakes manipulate enough people, enough of the time, to be a significant cause for public concern), and allow other researchers, academic, industrial (Google), and government (e.g., FBI) to explore the large breadth of follow-up questions that it sparks.

**Reviewer 2**: ii) It also remains unclear whether people would detect these stimuli as fake without being told that some stimuli were fake and others were not and then reflecting back on what they saw.

**Authors**: We agree with Reviewer 2. However, the point they make is a far more general one about the utility of self-report measures which applies to all of social science (i.e., that the act of measurement may perturb the system). It seems unfair to be held to a higher standard (or to be rejected because of an issue) that extends far beyond our own paper and applies to the broader field at large.

**Reviewer 2**: Using the current procedures at least, participants appear better than chance at detecting fake versus genuine materials (in Experiment 6). Yet, it could still be that reports of "faked" material are mostly just guesses held with little certainty. Such perceptions would be unlikely to have much effect on judgments. Based only on such judgments of "detection," it seemed like a bit of a stretch to say, for example, that "even detectable or imperfect Deepfakes can be used to manipulate a viewer's attitudes and intentions" (p. 6) or that "technological solutions designed to detect and flag Deepfaked content for viewers will not be enough" (p. 6). It is not clear how detectable the current Deepfakes really were (and presumably the amount of detectability would matter), and there is no evidence that flagging a video as fake, for example, before viewing it would be ineffective. It is entirely possible that such statements might ultimately hold true, but the current data do not justify them at this point.

**Authors**: Reviewer 2 argues that “it could still be that reports of "faked" material are really mostly just guesses held with little certainty.” Once again, this is a statement about the utility of explicit measures in social science and an issue that applies to most social science research reliant on self-reports.

Perhaps more importantly, the issue of detection (which the Reviewer seems to primarily focus on) was always and only a secondary issue in our paper. To be clear: our main goal was to investigate if (1) Deepfakes are good enough (not necessarily indistinguishable from genuine material) to manipulate enough people, enough of the time, to be a serious cause for public concern.

Our data repeatedly showed this to be true, and we ourselves were shocked to see how strong the effects were. This is why we ran multiple pre-registered exploratory and a separate confirmation study to ensure maximum reliability of our findings. It was only after we empirically demonstrated that (1) was true that we then turned to the separate and secondary questions of (2) can people detect when they are being exposed to Deepfakes, and (3) are Deepfakes similar to genuine content in their impact.

**Reviewer 2**: i) it seemed like a bit of a stretch to say, for example, that "even detectable or imperfect Deepfakes can be used to manipulate a viewer's attitudes and intentions" (p. 6)....

**Authors**: We are happy to revise this claim based on Reviewer 2’s suggestions. We could also cite research showing that even if people are consciously aware of biases, that is not enough to prevent them from being biased - which we expect to generalize to this area.

**Reviewer 2**: More importantly, however, I wondered whether a number of the claims based on the data are justified or, in some cases, how much a reader should make of what the data actually say. On some level, if the fakes are so well-done that they appear genuine, there is little reason why they would NOT be as effective as genuine material. As noted earlier, it may well be important to demonstrate that deepfakes can be so well-done that they are not detectable, but once that is demonstrated, there is little reason to question whether such content could have similar impact to (indistinguishable) genuine content. The additional questions about awareness and detection were seemingly intended to address whether increasing knowledge or labeling of material as fake might be effective in undermining the material's effects.

i) On some level, if the fakes are so well-done that they appear genuine, there is little reason why they would NOT be as effective as genuine material.

**Authors**: We strongly agree with Reviewer 2 on this point. However, the “fakes so-well-done” scenario is at one end of a spectrum, and not of immediate real-world applicability. And only applies after we have presented our study showing Deepfakes have such strong effects.

As we outlined above, our main aim was to answer a different question: (1) are “fakes good enough” (not necessarily “so well done”) that they manipulate enough people, enough of the time, to be a serious concern. Our data shows this to be true, which is the primary hypothesis test in this work (and never empirically shown before). To discount this result based on retrospective conjecture on “lack of surprisingness” would be a mistake - we should be interpreting it through the lense of its important societal implications.

We ourselves were shocked to see how strong the effects were - which is why we ran multiple pre-registered and a separate confirmation study to ensure reliability. It is only after 1) is true, and our data so strong, that a new question opens 2): why are the deepfakes and genuine material so similar, e.g., as the reviewer hypothesis’, is it because of the “so-well-done'' effect? We provide data to answer this. Regardless of 2), the original question 1) is of important and urgent concern and is why we believe the PNAS audience would find it immensely useful.

**As an aside to the Editor**, Reviewer 1 understands the real world implications of our research question (1) - e.g., they are in policy meetings where this is brought up. Reviewer 2 is answering a separate and more academic question (2), that deserves to be studied further - but it does not have the high impact, real-world implications of (1).

**Reviewer 2**: In the case of awareness, similar to studies of awareness of bias, researchers often examine awareness not through asking research participants whether they are aware that bias can exist (in everyday life). Rather, they often manipulate awareness by making a potential for bias salient in some way prior to encountering the stimuli (sometimes rather indirectly, sometimes more directly). In the current case, making people aware of the potential for deepfakes might involve something like having people read an article describing deepfaked audio or video (or not) prior to exposure to the target stimuli. But that is not at all the kind of awareness examined in the current work. Rather, awareness was assessed after participants encountered the stimuli and after, in many of the studies, people were told that the stimuli they saw was faked.

i) researchers often examine awareness not through asking research participants whether they are aware that bias can exist (in everyday life). Rather, they often manipulate awareness by making a potential for bias salient in some way prior to encountering the stimuli (sometimes rather indirectly, sometimes more directly).

**Authors**: We appreciate Reviewer 2’s comment that there are alternative ways of measuring awareness. However, we designed our studies intentionally in order to be as closely aligned with a real world scenario as possible (people being exposed to a Deepfake and only then being made aware of this fact). Utilizing a different methodology isn’t wrong, just different, and useful information can still be obtained - as was the case for us. Future studies can indeed pursue this. However, it should be noted that the awareness/detection questions are very much less consequential to our main research question (1) above (Deepfakes are incredibly harmful right now), and more an academic question of (2) (why are Deepfakes similar genuine content).

**Reviewer 2**: In the end, the manuscript reports effects of high applied significance, but the conceptual contributions remain somewhat unclear. The potential links between awareness, detection, and influence of deepfakes are not yet well-understood, but perhaps those are not really the key points for the article to make. At a minimum, however, it would be useful for the manuscript to be a bit more circumspect about what can be assumed or claimed on the basis of the current awareness and detection data.

**Authors**: We appreciate Reviewer 2’s comments and summary. With the clarification of the incorrect understanding of detection, and the awareness methodology, we hope the conceptual contributions of our work are now better illuminated.

The reviewer is correct that the detection/awareness are not the key points of the article, rather, as mentioned above, our first priority is (1) showing that Deepfakes are an imminent societal danger to enough people, enough of the time. Deepfakes being so similar to genuine material surprised us, and led to the secondary question of (2) “why is the effect size the similar between deepfakes and genuine material” (are people not aware, are people not detecting them, etc). Regardless of 2), Deepfakes are incredibly harmful (as we note to the Editor: the FBI speculates Deepfakes of novel individuals - like those studied in our paper - will be used for mass manipulation purposes within the next 12-18 months; https://www.ic3.gov/Media/News/2021/210310-2.pdf). This point is of high relevance to the broad readership of PNAS.

If we interpret Reviewer 2 correctly, they are not suggesting an outright rejection, but rather more tempered conclusions with respect to awareness/detection, which we would happily accommodate.