Manuscript No. XGE-2021-3737  
Deepfaked Online Content is Highly Effective in Manipulating Attitudes & Intentions  
Journal of Experimental Psychology: General  
   
Dear Dr. Hughes,  
   
I have received reviews of the manuscript entitled Deepfaked Online Content is Highly Effective in Manipulating Attitudes & Intentions (XGE-2021-3737) that you recently submitted to *Journal of Experimental Psychology: General*. I was fortunate to receive comments and evaluations from individuals who are very knowledgeable and highly respected experts in the topical area you are investigating. As you will see when you read their critiques, the reviewers have offered many detailed points and constructive suggestions centered on improving the current paper.  
   
I read the manuscript prior to receiving these reviews in order to gain an independent perspective on the paper, and then again with the reviews in hand. In the end, there turned out to be a considerable level of consensus among the majority of us with respect to the perceived strengths and limitations of the current paper. All of us found several aspects of the work appealing, namely the methodological novelty and timeliness of the topic.

**Authors**: We thank the Editor and Reviewers for their kind words, as well as constructive and thorough feedback. It helped us when carrying out a major revision of our paper and resulted in a far stronger contribution.  
   
**Editor**: At the same time, however, the reviewers raised some concerns that prevented them from recommending acceptance of the paper in its current form. I share some of these same concerns and in fact was shocked (pleasantly) that all reviewers identified issues that I had identified in my own reading. I weighed the critiques against the enthusiasm behind a potential contribution and ultimately decided that I would like to encourage you to submit a revision. However, such a revision would need to include new data and there is no guarantee that your manuscript will be published.

Because the reviewers' comments are clearly expressed, I will not reiterate all of the issues that they have raised. It is rare for me to say this, but I agree with essentially all the points they have raised and believe they need to be addressed. Below I discuss the major issues that I identified as well as points of convergence that prevent me from recommending at present publication in Journal of Experimental Psychology: General.

The biggest issue, in my view, is trying to figure out what the major contribution of the work is. Without getting too deep into philosophical questions, I was wondering what does this set of findings really tell us about the effect of falsified stimuli on our perceptions? Deepfakes may misrepresent their subjects, but they are certainly real content insofar as we can see them and hear them. Why would we expect them NOT to affect perceptions? In many ways, I felt these studies were conceptually similar to showing people the Mona Lisa, as well as a perfect forgery of the Mona Lisa, and asking them to respond to both. Would we not expect people to have a similar reaction between an authentic Mona Lisa and an exact forgery? What about telling participants a plausibly believable lie? Of course, we would expect people to respond to the content of the lie in the same way as a truthful statement with the same content.

**Authors**: The Editor and Reviewers ask about our work’s main contribution. For instance, the Editor notes: “*these studies were conceptually similar to showing people the Mona* *Lisa, as well as a* ***perfect forgery*** *of the Mona Lisa, and asking them to respond to both. Would we not expect people to have a similar reaction between an authentic Mona Lisa and* ***an exact forgery***?”. The Reviewers echo similar sentiments.

The above assumption boils down to the following: if (a) Deepfakes are *perfect replicas* of authentic content then (b) they should influence perceptions to the same extent as such content.

We agree with the *logic* behind this point. However, the premise upon which it is founded is problematic. The vast majority of Deepfakes are *not* perfect replicas of authentic content but imperfect copies that vary drastically in their respective quality and believability. Most, including our own, contain audio and visual artefacts. For several real-world examples see <https://www.youtube.com/watch?v=ZJrffEfCMrs>.

These visual and auditory artefacts represent perceptual cues that signal to the viewer that what they are watching or listening to has been tampered with, artificially constructed, or otherwise edited or modified. These cues *should* lead viewers to question what is being communicated to them, which in turn should undermine the impact of that information on their thoughts, feelings, and actions. Our findings consistently show that this was not the case. Deepfakes of the target quickly and powerfully shifted attitudes and intentions despite the presence of such cues. More interestingly, even people who recognize that what they are watching is fabricated or manipulated (‘Deepfake detectors’) and who were aware of this technology before being exposed to it, still fell prey to its influence.

Reflecting back on our original submission, we could have done a better job of communicating this point (i.e., that the vast majority of online Deepfakes are NOT perfect replicas of authentic content; that they are imperfect informational sources that contain invalidity cues which should trigger the perceptual system to question what one is seeing or hearing, to reject that information, and minimize the impact of that information on ones thoughts and feelings). We have now revised the manuscript to better communicate these ideas (see changes on p.14, 39-43).

Finally, we now do a better job of contextualizing our work in the wider Deepfaking literature, explaining how it related to what has come before, and extends upon early work in important new ways (e.g., see p.11-13; 38-39). Specifically, we now explain why its so important to examine how Deepfakes can be used to target members of the general public as was the case in our studies:

“It may be tempting to assume that Deepfaking only poses a problem for the most visible members of society (e.g., politicians, industry leaders, celebrities) given that first-generation Deepfakes have largely centered on these individuals. However, the rapid spread of open-access and freely available Deepfaking tools undermines any such assumption. The technology is readily accessible to the general public and poses a risk to anyone who has ever posted content of themselves online. Such content can easily be scraped from social media, fed as a training dataset to one of these apps, and used to generate a Deepfake of that individual.

This is already a lived reality for school children (Morales, 2021) and journalists (Ayyub, 2018) being cyberbullied via Deepfakes, women facing reputational and emotional harm from Deepfaked revenge porn (Hao, 2021), and CEOs subject to identity theft and fraud (Stupp, 2020). The rise of Deepfakes brings with it an increased risk of citizen impersonation and online abuse, especially towards the most vulnerable members of society (e.g., a cybercriminal might pose as a family member in urgent need in order to extract money from elderly relatives online; EU Commission, 2018). As we noted at the outset, international media outlets are already being targeted, and social media platforms infected, by disinformation networks comprised of AI-generated personas of ‘normal’ people.

With the above in mind, we set out to answer three questions. First, how effective are Deepfakes in biasing our implicit and explicit attitudes towards people we are meeting for the first time? Despite a widespread belief that Deepfakes can shift attitudes, only one study has empirically examined this issue so far, and it did so while focusing on a well-known individual (politician), explicit (but not implicit) attitudes, and on a highly specific topic (Christian political statements) (Dobber et al., 2021). It remains to be seen if a single brief exposure to Deepfaked online content is capable of manipulating our implicit and explicit first impressions of others. Second, several studies have examined the viral side of Deepfakes (i.e., their intentions to share content of known individuals with others on social media platforms). We were curious to know how readily people would share Deepfaked content of novel individuals. Third, we examined how many people are aware that Deepfaking is possible (awareness), and if they could detect whether they had been exposed to one (detection). We wanted to know if awareness and detection would serve to immunize them from being influenced by Deepfakes.”

**Editor**: The major contribution of the paper, in my opinion, therefore, rests on Study 6. What I was seeking all along, as I read through Experiments 1-5, was a study where participants explicitly knew that (or were made aware that) the content of a Deepfake was false, but showed comparable perceptions to genuine content. Experiment 6 achieves this to some degree. The authors also note “Deepfake awareness and detection were also probed in Experiment 3, but I did not see any analyses on this point—they should certainly be included.

**Authors**: The Editor is correct that a question related to Deepfake detection was included in two of our exploratory studies (Experiments 2 and 4). However, this was a *directed* question: participants were first informed what a Deepfaked was, told them that they had just been exposed to one, and then asked them to indicate whether they had been aware of this fact while watching the content. They then responded using an open-ended text box.

There are three reasons we omitted analyses on this question from the manuscript. First, and upon reflection, we realized that the question did not ask participants *if* they had encountered a Deepfake. Rather it told them that they had done so: both in the authentic and Deepfaked conditions. This style of phrasing may have influenced how participants responded and does not reflect detection *per se*. Second, the open-ended responses may have introduced a degree of subjectivity into the scoring of these responses. Third, including these analyses may encourage readers to make a comparison between the detection questions from the exploratory and confirmatory studies which would be inappropriate (given that the former and latter appear to be addressing different issues).

Critically, we took each of these issues into mind and addressed them in our high-powered, confirmatory study (see Experiment 7). In that study we (a) asked rather than told participants if they had encountered a Deepfake during the study, and (b) used a close ended Yes/No question. We continue to believe that the exploratory detection analyses should be omitted from the manuscript. That said, we are happy to add them in if the Editor still sees merit in their inclusion.

**Editor**: Even with the findings of Experiment 6, I wondered, how is this different from watching a movie where you know the story is pure fiction? We know that as long as the story is compelling, it is certain to produce an emotional response, which is what I feel is shown in this study. I was left wanting more data to demonstrate a contribution to the psychological literature. Reviewer 1 shares my enthusiasm for Experiment 6 and made me think that this study could be the jumping off point for a future revision.

Reviewer 2 also offers a useful suggestion for a design that would manipulate Deepfake detection awareness rather than simply measuring it. Perhaps more concerning, Reviewer 3 questions whether the measures of Deepfake detection in Experiment 6 are adequate for capturing whether participants really were able to distinguish genuine content from deepfaked content.

**Authors**: The Editor asks here: “*how is this different from watching a movie where you know the story is pure fiction?*”. It’s worth bearing in mind that participants did *not* know which condition they had been assigned to at any point during the study. As far as they were concerned the content they were watching was authentic in nature, and they were only informed about that the content *may* have been Deepfaked at the very end of the study.

We also agree with the Editor’s second point that “*as long as the story is compelling, it is certain to produce an emotional response*”. This is one of the dangers of Deepfakes – by conveying compelling information they can bias what people think and feel, despite clear audio and visual cues signaling that what one is watching is likely to be fake. We now include new material in the General Discussion unpacking this issue (see p.42). We have also included a new section discussing open questions and future directions for psychological research on Deepfakes (see p.43-50)

We have also acknowledge the limitations of our approach to Deepfake detection (see p.51) and added Reviewer 2’s suggestion for a design that would manipulate Deepfake detection awareness rather than simply measuring it (see our response to Reviewer 2 and changes on p.51) and responded to Reviewer 3’s questions about Deepfake detection in Experiment 6 (see our responses to Reviewer 3 and new material on p.X).

**Editor**: Other reviewers questioned the contribution of the work as well, with R1 noting that the work feels a “bit tautological; a piece of content becomes a "deepfake" when it has been manipulated so seamlessly that it appears as if it were genuine content.” R2 echoes this point almost identically, stating, “Deepfakes, by definition, refer to videos that are believable/seemingly authentic fakes that can deceive viewers.”  This again raises the question that I posed above, which is what is the contribution of knowing that something that has been designed to replicate genuine content produces similar psychological responses to actual genuine content? All reviewers have additional insightful comments about strength of contribution. I urge the authors to consider these comments carefully.

**Authors**: We respond fully to Reviewer 1 and 2’s point about tautology and believability below. But to briefly preface those arguments, it is not accurate to say that “*a piece of content becomes a deepfake when it has been manipulated so seamlessly that it appears as if it were genuine content*”. Or that “*Deepfakes, by definition, refer to videos that are believable/seemingly authentic fakes that can deceive viewers*”.

As we now state in the revised manuscript, AI-generated media can be divided into two broad categories. The first involves the use of AI to generate videos, images, audio, and text of individuals who do *not* exist (e.g., see <https://this-person-does-not-exist.com/en>, <https://www.resemble.ai/>, and <https://openai.com/blog/openai-api/>). This category is known as synthetic media.

The second involves the use of AI to generate or manipulate videos, images, audio, and text of individuals who *do* exist. It is this class of media that is commonly referred to as ‘Deepfakes’. Deepfakes can involve mimicking a person’s image (photo), actions (video), voice (audio) or writing style (text). As such, Deepfakes are defined by two properties: (a) how they are created (using AI), and (b) what type of content is being operated upon by that AI method (i.e., content involving *existing* individuals).

Unlike the reviewers, we don’t believe it’s accurate to define Deepfakes using concepts such as believability, authenticity, or deception. Any type of media (Deepfaked or not) can – in principle - vary along these and other dimensions (e.g., a real photo, painting, CGI, or any other type of media can vary in its believability or perceived authentic). Similarly, any type of media can be used to deceive – not just Deepfakes. And not all Deepfakes are used for deceptive purposes (see p.5 for relevant examples).

In short, how Deepfakes have been defined by the reviewers, and the premise that they are perfect replicas of genuine content (and thus why wouldn’t they produce similar outcomes) are both problematic. Once one sets these premises to the side, the novel contribution of our work becomes clear (i.e., that disinformation about others communicated via Deepfakes influence attitudes and intentions despite the presence of invalidity cues signaling that what one is watching or hearing is fabricated/manipulated). We now articulated these points more clearly in the revised manuscript.

**Editor**: One final note on contribution is that I did not feel that Experiments 1a and 1b provided value given they simply tell us that genuine content can affect attitudes (which we already know). Given that these studies do not seem to be used to compare perceptions of Deepfakes to perceptions of genuine content in subsequent studies, they do not add much to the paper.

**Authors**: In-line with the Editors suggestion we have removed Experiments 1-2 from the revised manuscript and now focus exclusively on those studies that deal with Deepfaked content (see revised manuscript and footnote 8 on p.15).

**Editor**: A separate issue that came up multiple times was simply that writing of the methods and results sections is not as clear as it could be, and there are inconsistencies with the OSF link as well. Reviewer 2 sums this issue up well by stating, “The format for the methods section adopted here is unconventional and lacks clarity.” As one example, for Experiment 2, I wasn’t clear on what the comparison was between Deepfake vs. genuine content perceptions that supports the statement, “They also produced attitudes that were just as strong as those established by authentic content.” I assumed that for this experiment (and for Exp2-6 generally) this simply means that within a particular study, perceptions of genuine content were compared to perceptions of Deepfake content, but when I went to the OSF site for Exp2 and looked at “stimuli” I only saw “genuine” videos and was further confused.  In general, I felt that the manuscript could spend more time walking readers through the analyses and procedures, and other reviewers echo this point. I also think more could be done with the existing data, particularly with regard to tracking convergence or divergence between explicit and implicit attitudes.

**Author**: We thank the Editor and Reviewers for pointing this out to us. We have extensively revised the paper to better communicate our methods, analyses, and conclusions (see updated manuscript). We have also modified our OSF project page to make it easier for readers to navigate and interact with our various materials (see <https://osf.io/f6ajb/>).

**Editor**: A final point of convergence for me and the reviewers is lack of theoretical development. Reviewer 2 very helpfully suggests exactly what a more robust theory section would look like, including highlighting the findings of Exp6 to discuss why Deepfakes would influence attitudes even if their inauthenticity was detected. Reviewer 3 also suggests ways of enhancing the literature review on deepfakes and very helpfully identifies existing papers that should be incorporated. A revision would need to go beyond simply adding these papers to the introduction and would need to genuinely grapple with the existing literature to both formulate hypotheses and make a convincing case that the present work represents a novel contribution.

**Authors**: We thank Reviewer 2 and 3 for their constructive feedback. In line with Reviewer 2’s suggestion, we have added new material on the theoretical implications of our work (see p.39-43). In line with Reviewer 3’s suggestions, we have deepened our consideration of the literature surrounding Deepfakes and incorporated this into both the Introduction and the General Discussion (see p.6-11; 39-43).

*Note*: we formulate potential explanations for our findings based on the wider literature and existing theory. But we restricted these to the Discussion section, and did not generate hypotheses, as to do so would constitute a case of hypothesizing after results are known (HARKing). We believe the major re-write of our paper based on the Editor and Reviewers’ comments has resulted in a far stronger contribution.

**Reviewer 1**: This was a bit of a puzzling review for me to write, as I saw a lot of potential for this line of work to produce some compelling and relevant findings, but disagreed pretty strongly with the notion that the present submission merits acceptance at a journal like JEP: General. In particular, I seem to disagree with the authors about the need to investigate the "open question" about whether Deepfaked material produces a similar psychological effect as genuine material.

My concern here is a bit tautological; a piece of content becomes a "deepfake" when it has been manipulated so seamlessly that it appears as if it were genuine content. The definition of "deepfake" from merriam-webster.com seems to agree with this point of view, referring to content that "has been edited using an algorithm to replace the person in the original video with someone else (especially a public figure) in a way that makes [it] look authentic." (emphasis added). I feel as if a video were so obviously manipulated such that it produced effects that differed from genuine content, well then that wouldn't be a deepfake! For these reasons, I did not find the results of Experiments 1-5 to be very surprising or of much interest, other than in proving that current Deepfaking technology is quite impressive and providing a proof of concept for using such technology in studies about attitude formation.

**Authors**: We thank Reviewer 1 for their feedback. The first comment centers on the following: “*a piece of content becomes a ‘deepfake’ when it has been manipulated so seamlessly that it appears as if it were genuine content*”, and argues that this idea is in line with the definition of a "deepfake" on merriam-webster.com.

We see several issues with both the idea and definition. We will respond to each below.

***Issues with the definition of Deepfakes***. The definition Reviewer 1 offers is *not* how the term Deepfake is typically conceptualized or defined within the scientific literature (in much the same way that many technical definitions in psychological and the physical sciences differ from lay terms used in online dictionaries).

Deepfakes are better conceptualized as a sub-type of a much larger class of media known as AI generated media (i.e., content which is generated or manipulated via artificial intelligence techniques such as Generative Adversarial Networks [GANs] and related methods). AI-generated media can be divided into two broad categories.

The first involves the use of AI to generate videos, images, audio, and text of individuals who do not exist (e.g., see <https://this-person-does-not-exist.com/en>, <https://www.resemble.ai/>, and <https://openai.com/blog/openai-api/>). This is known as synthetic media.

The second involves the use of AI to generate or manipulate videos, images, audio, and text of individuals who *do* exist. It is this class of media that is commonly referred to as ‘Deepfakes’ and which can involve replacing one person with another (e.g., face swapping) OR fabricating novel content of the same person (e.g., using AI to mimic the writing style or voice of an existing individual).

Thus Deepfakes are defined based on two properties: (a) how they created or manipulated (via AI methods), and (b) what type of content is being operated upon by that AI method (i.e., content involving *existing* individuals). Unlike the definition offered above (which exclusively center on videos) images, text, and audio can also be Deepfaked, and don’t simply involve replacing one individual with another. Perhaps most importantly, as we outlined in our response to the Editor, we don’t believe that Deepfakes should be defined in terms of properties such as believability, authenticity, or deception.

We discuss this latter point in more detail below (also see footnote 1 on p.8 of the revised manuscript). But for now it’s important to note that the definition of Deepfakes that was offered is problematic for us on multiple grounds.

***Issue with equating Deepfakes with certain media properties or usage***. When Reviewer 1 says “*a piece of content becomes a ‘deepfake’ when it has been manipulated so seamlessly that it appears as if it were genuine content*” they are defining a Deepfake in terms of its *properties* (e.g., believability) and its *purpose* (e.g., to deceive others). Yet these properties and purposes are not unique to Deepfakes at all – they are shared by any type of media. Put another way, any video, text, audio, or photo can be used with a particular purpose (to deceive viewers) and can vary along dimensions such as its perceived believability or authenticity. Thus the idea that “seamless manipulation” and “appears genuine” define Deepfakes is problematic.

It’s worth noting here that Deepfakes are typically not “*manipulated so seamlessly*” to appear perfectly genuine (i.e., they vary drastically in their quality and thus believability; see <https://www.youtube.com/watch?v=ZJrffEfCMrs>). The idea that Deepfakes are seamless replicas of genuine content that are designed to perfectly deceive viewers fails to reflect the technology as it is used today or the vast majority of Deepfakes available online.

Likewise, Deepfakes are not always used to deceive (see p.5 of the revised manuscript for real-world examples).

In short, we don’t believe that Deepfakes should be equated with, or define in terms of, certain properties (perfect replicas of authentic content) or purposes (deception) seeing as these very same concepts apply to any type of media, and don’t apply to all types of Deepfakes. Indeed, it may be that people encounter Deepfakes that are clearly manipulated, not perfectly believable, and which still influence their thoughts, feelings, and actions (e.g., for one recent attempt to do so with President Zelensky see <https://www.youtube.com/watch?v=X17yrEV5sl4>). According to the webster.com definition and Reviewer 1’s comment, these would not qualify as Deepfakes.

Finally, when Deepfakes are viewed in this way, the contribution of our work becomes clear: we provide empirical evidence that even imperfect Deepfakes that are clearly not perfect replicas of genuine content still strongly bias people’s implicit attitudes and intentions, even people are aware of Deepfaking and have detected that they have just been exposed to one. We discuss these and related ideas in the revised manuscript.

**Reviewer 1**: I actually think it would be a potentially nice contribution to write up Experiments 1-5 as a more general methodological piece introducing psychologists to deepfake technology as a way of creating more control over experimental stimuli (e.g., recording one video about positive information and then deepfaking the negative information condition may produce more similar stimuli than just recording two separate videos).

**Authors**: This is a really nice idea. We have added new material unpacking it in the General Discussion (see footnote 16 on p.45).

**Reviewer 1**: Experiment 6 was a notable exception in that I did find it quite interesting that the effects of the deepfaked content were similar for people who did vs. did not accurately label the video as a deepfake. I could see this being a good Study 1 for a larger investigation of this effect. One concern I had about the present study is that participants are being exposed to a novel target, so they may be more likely to doubt the presence of a deepfake (i.e., why would someone go through the trouble of Deepfaking a video for a person I don't know anything about?). I think some interesting follow-ups here would be to see if similar effects emerge for well-known targets (e.g., a deepfake of Joe Biden espousing his love of communism) and then to further investigate if the effectiveness of a deepfake is moderated by pre-existing attitudes towards Biden. It may then be of interest to see if similar manipulations known to be effective against misinformation in other forms (e.g., Pennycook et al., 2020; Pennycook et al., 2021) are also effective against deepfakes.

**Authors**: Reviewer 1 asks “*why would someone go through the trouble of Deepfaking a video for a person I don't know anything about*?”. There are many reasons why one would do this in real-life.

They may want to construct a false identity for themselves as a journalist in order to manipulate international newspapers into publishing certain content (for a recent real-world example see <https://medium.com/general_knowledge/deepfake-journalists-fake-news-e8c3c3af70c9>). They may want to fabricate a political candidate and use them to either support or discredit a political party or message before or during an election (for examples of what this might look like see: Dobber, T., Metoui, N., Trilling, D., Helberger, N., & de Vreese, C. (2021). Do (microtargeted) deepfakes have real effects on political attitudes?. The International Journal of Press/Politics, 26(1), 69-91). An angry ex-partner, disgruntled work colleague, or malicious actor could scrape a victim’s image from social media and insert it into a pornographic video, thus biasing how the victim is perceived by others upon the release of that content (for real-world examples of this see <https://www.technologyreview.com/2021/02/12/1018222/deepfake-revenge-porn-coming-ban/>). In short there are many meaningful ways in which Deepfakes of novel individuals are already being used in everyday life. Studying and understanding the psychological impacts of such content is therefore a worthwhile endeavor. We have added new material outlining how Deepfakes of novel individuals are being used to spread disinformation and to target members of the general public p.4-5 and 11-12).

Reviewer 1 makes a nice point about moderation of Deepfaking effects by prior knowledge of the target (e.g., would similar effects emerge for known [political, celebrity] compared to non-known individuals?). This was actually one of the very first ideas we had in our project. However, we were not granted ethical approval by our IRB to explore this issue. We have nevertheless included additional material in the General Discussion highlighting this as an important direction for future research (see p.43-45). The same goes for the Reviewer’s other suggestion (i.e., whether manipulations known to be effective against other forms of misinformation are also effective against deepfakes) (see new material on p.48-50).

**Reviewer 1**: All of this is to say that I believe this to be a promising line of research, but that the evidence concerning the psychological factors related to processing deepfaked content are quite underdeveloped.

**Authors**: In line with Reviewer 1’s suggestions, we have elaborated on the psychological factors related to Deepfaked content in our revised manuscript (see new material on p.39-43) and also highlight open questions and future directions for psychological research on Deepfaking (see new material on p.43-50).

**Reviewer 2**: In this paper, the authors aim to examine the psychological impact of deepfakes, and find that deepfakes are as effective as genuine content in influencing people's implicit and explicit attitudes and behavioral intentions. I was excited to review research on this topic as Deepfaking is a relatively new phenomenon that warrants attention and understanding its psychological impact can have important implications. The experiments were simple and interesting, and I appreciated the authors' attempts to create realistic experimental paradigms that accurately capture the phenomenon of study and explore the impact of both audio and video stimuli. I also appreciated the authors embracing Open Science and sharing their materials and data.

**Authors**: We thank the Reviewer for both their kind words and constructive feedback.

**Reviewer 2**: However, the paper is not without limitations. Below are my main concerns about the paper:  
  
1. *Scope of the contribution*: As noted earlier, I believe understanding the psychology of deepfakes is important. Given the importance of the topic, it is unclear how the paper contributes to deepening our understanding of the psychological impact of deepfakes. Let me provide some reasoning here. Deepfakes, by definition, refer to videos that are believable/seemingly authentic fakes that can deceive viewers. Indeed, research on deepfakes refer to these as videos created by artificial intelligence/machine learning (AI/ML) applications that "merge, combine, replace, and superimpose images and video clips onto a video, creating a fake video that appears authentic" (Maras & Alexandrou, 2018).

Furthermore, research on deepfake detection begins with the premise that deepfake videos are realistic and believable and have the potential to cause widespread societal harm due to the very realistic nature of these videos (e.g., Güera & Delp, 2018). Given that being a realistic and believable fake is the very basic quality of a deepfake, the current findings suggesting that deepfakes are, in fact, as good as genuine content in being believable and influencing people's attitudes is somewhat underwhelming. The paper's findings are all quite straightforward and essentially confirm what we know about deepfakes already - they are good at influencing people's attitudes. The effectiveness of deepfakes is precisely the reason why we see huge efforts from researchers, technology companies, and governments across the world to detect deepfakes. I'd like to defer to the editor here, but I'm worried if these findings constitute a big enough contribution for a top journal like JEP:G.

**Authors**: Reviewer 2 makes two points here: the first centers on the definition of Deepfakes while the second centers on the scope of our contribution. We respond to these two points below (also see our response to the Editor and Reviewer 1).

***Definition of Deepfakes***. Reviewer 2 states that “*a realistic and believable fake is the very basic quality of a deepfake*” and that “*Deepfakes, by definition, refer to videos that are believable/seemingly authentic fakes that can deceive viewers*”. On the one hand, we agree with the Reviewer. Certain Deepfakes are videos, certain Deepfakes are believable/seemingly authentic, and certain Deepfakes are used to deceive others.

On the other hand, this definition is not quite right as there are many Deepfakes which do not satisfy these three criteria. For instance, Deepfaking is not just about creating fabricated videos. It’s an AI-based technique for generating or altering *any* content involving existing individuals. The technique can be applied to any type of media insofar as one can Deepfake images, text, and even audio (we now discus this point on p.5 of the revised manuscript).

Second, we believe it’s problematic to equate Deepfakes with specific media properties such as believability or perceived authenticity. When Reviewer 2 says “*Deepfakes, by definition, refer to videos that are believable/seemingly authentic fakes that can deceive viewers*” they are defining Deepfakes in terms of certain *properties* (e.g., believability) and certain *purposes* (e.g., to deceive others). Yet these properties and purposes are not unique to Deepfakes –any type of video, text, audio, or photo can be used to deceive viewers and can vary in their perceived believability or authenticity. The online information eco-system also contains many Deepfakes that vary drastically in their believability/authenticity (e.g., see <https://www.youtube.com/watch?v=ZJrffEfCMrs>).

Finally, Deepfakes should not be defined in terms of a particular use (e.g., to deceive people). There are many cases where Deepfakes are not being used to deceive (e.g., [Disney](https://screenrant.com/book-boba-fett-luke-skywalker-cgi-hamill-improved-explained/) is exploring the use of Deepfakes for creating extras in its movies and tv series, the [fashion industry](https://www.forbes.com/sites/forbestechcouncil/2019/05/21/gans-and-deepfakes-could-revolutionize-the-fashion-industry/) is using the technology to allow consumers to see themselves in the latest clothing lines, or to protect the identities of persecuted minorities in online and [documentary](https://theconversation.com/deepfake-technology-unlocks-real-stories-of-lgbtq-persecution-in-welcome-to-chechnya-144053) settings).

To conclude, we agree with the definition offered by Reviewer 2 insofar as its partially accurate (e.g., Deepfakes *can* be video that *can* be realistic, *can* be believable and *can* be used to deceive others). But we believe it’s also incomplete (conflates Deepfaking as a technique with one particular type of media [videos]) and is overly restrictive (only focuses on a sub-set of Deepfaked media – namely – videos that are highly realistic, believable, and which set out to deceive). For this reason, we offer a more general definition in footnote 1 on p.5.

***Scope of our contribution***. Reviewer 2 states the following: “*being a realistic and believable fake is the very basic quality of a deepfake”* and that *“…the current findings suggesting that deepfakes are, in fact, as good as genuine content in being believable and influencing people's attitudes is somewhat underwhelming*”.

At its core, this point boils down to the following: (a) Deepfakes - by definition - are realistic and believable fakes of authentic content, and therefore (b) why would we *not* expect them to produce similar effects as authentic content?

As we note above, this premise (that Deepfakes are perfect replicas of authentic content) does not reflect the reality of the technology or the type of Deepfakes currently present online. “Being a realistic and believable fake” is ***not*** “a very basic quality of a Deepfake”. Rather realism and believability are two continua along which Deepfakes vary, from those that are lower on these dimensions (e.g., see <https://www.youtube.com/watch?v=ZJrffEfCMrs>) to those that are higher (e.g., <https://www.youtube.com/watch?v=oxXpB9pSETo>).

This means that the vast majority of Deepfakes actually contain invalidity cues: signals that what the viewer is seeing or hearing is not a perfect replica of reality but rather a contaminated information source that may have been edited, constructed, or otherwise manipulated.

We expected that these cues would be factored into the viewer’s reasoning about the information we provided and moderate its impact on their thoughts, feelings, and actions. Yet we found that this was *not* the case: even imperfect Deepfakes containing signs of manipulation quickly and powerfully shifted (implicit) attitudes and intentions. This was still true when people identified that the content they watched or listened to had been tampered with (‘Deepfake detectors’). Thus even imperfect misinformation sources bias people to similar extents as authentic informational sources.

To conclude, when one sets the problematic idea that Deepfakes are perfect replicas of authentic content to the side, and instead engages with the reality of the technology, then the novel contribution of our work becomes clear (i.e., that disinformation communicated via Deepfakes can quickly and effectively influence the recipient, despite clear signs that what they are watching or listening to has been fabricated or manipulated).

We did not do a good enough job of communicating these points in our original submission. We have revised our manuscript to better address these issues and situate our contribution in the wider literature (see the many changes made through the revised manuscript).

**Reviewer 2**: *Theory development*: The paper focuses on research questions that are empirically driven rather than theory-driven. This is not a concern in and of itself, but is particularly problematic for top psychology journals like JEP:G where readers tend to expect theoretical insights that can deepen our understanding of the phenomenon of study. I believe there is a missed opportunity for theorizing here: the authors can examine why deepfakes are so effective in influencing people's attitudes and explore the psychological mechanisms driving this effect. Furthermore, the authors can also examine and theorize about whether people are more likely to make certain types of moral judgments versus others about people based on deepfakes. A very interesting, yet underexplored, aspect of the current paper is the 'why' behind deepfakes' impact on attitudes and behavioral intentions even when people are aware or can detect deepfakes (see point#3). All these directions could lead to better theorizing and a bigger theoretical contribution.

**Authors**: We appreciate the Reviewer’s point. As they correctly identify, we started this project with an empirical rather than theoretical agenda (e.g., to examine if disinformation communicated via Deepfakes could shift [implicit] attitudes and intentions).

That said, we see that we could have done more to articulate the “*why behind Deepfakes*” (i.e., to “*examine why deepfakes are so effective in influencing people's attitudes and explore the psychological mechanisms driving this effect*”). “*Examine and theorize about whether people are more likely to make certain types of moral judgments versus others about people based on deepfakes*.” And to hypothesize about why Deepfakes shift “*attitudes and intentions even when people are aware or can detect deepfakes*”.

We took this comment seriously and revised the General Discussion to better address these questions. *Note*: we restricted this material to the Discussion because as we did not want to give the reader the mistaken impression that we generated hypotheses based on this new material (to do so would be to fall prey to HARKing), or that this material was the starting point for our work (which it was not) (see changes on p.39-50).

**Reviewer 2**: *Deepfake awareness/detection* and its impact on attitudes and behavioral intentions is an aspect of the paper that has the potential to address important questions in this area. However, this aspect of the paper is largely underexplored. From a theoretical standpoint, there is virtually no theorizing about when and why deepfake awareness/detection can influence attitudes as this is just listed as a question (p. 6).

**Authors**: see the above comment and changes made on p.39-43.  
  
**Reviewer 2**: The pre-registration for Experiment 6 lists research questions that pertain to the effectiveness of deepfakes in establishing first impressions, but first impressions are not discussed in the current theorizing. The narrative front end of the paper largely focuses on general implicit and explicit attitudes. Deepfakes have largely become notorious for creating inauthentic content of well-known people (e.g., world leaders) about whom people already might have formed prior impressions. It would be theoretically interesting to explore the effectiveness of deepfakes in shaping attitudes about well-known people versus strangers.

**Authors**: In line with the Reviewer’s suggestion, we have added new content to the General Discussion that explores open questions and future directions for research on Deepfaking and attitudes. Here we discuss the need for work examining attitude formation and change for both novel and known individuals, the longevity and durability of Deepfake-induced attitudes, their malleability, and factors likely to moderate when Deepfakes have a maximal vs. minimal impact on attitudes (see changes on p.43-45).  
  
**Reviewer 2**: I expected more of a discussion about how deepfakes affect implicit versus explicit attitudes. Did the authors expect differences between implicit versus explicit attitudes, but didn't find any? Some more context about how readers should think about these results would be important.

**Authors**: Cards on the table: we included implicit and explicit attitudinal measures to (a) examine if Deepfakes shifted both attitude types (i.e., and thus show that our arguments generalize across different measures), and (b) we had no *a priori* theoretical assumptions about the capacity of Deepfakes to differentially impact implicit vs. explicit attitudes.

That said, and in-line with the Reviewer’s suggestions, we have included new material in the General Discussion acknowledging outlining potential future directions for research on this topic (see changes on p.49).  
  
**Reviewer 2**: Another missed opportunity for theoretical contribution pertains to the differences between audio vs. video Deepfaked content. It was interesting to see similar results across audio and video stimuli. Are there theoretical reasons why we might expect similar results across the two stimuli? What makes deepfakes equally effective in both audio and video forms?

**Authors**: In line with the Reviewer’s suggestions, we have now added new material discussing audio vs video Deepfakes, and the potential reasons why similar outcomes occurs for both media types (see changes on p.42 and footnote 14)  
  
**Reviewer 2**: From the description of experimental designs in the paper and in the supplemental materials in the OSF, it is unclear whether the designs for Experiments 2 - 6 are a 2 (positive vs. negative) x 2(genuine vs. deepfake) between subjects design, and if the authors are predicting two main effects only, two main effects and an interaction, or any other combination of main effects and interactions. The lack of theory also makes this murkier, and difficult to interpret the design choice and findings.

**Authors**: Reviewer 2 is correct in that we adopted a 2(*Informational Content*; positive vs. negative) x 2 (*Content Type*: genuine vs. Deepfaked) design in our exploratory studies. However, in our pre-registration document we stated that we would only focus on the main effects of Information Content and Content Type and not the interaction between the two. We have clarified this on p.22. We have also conducted a non-pre-registered re-analysis of our exploratory findings based on the Reviewers comment. Briefly, the same general pattern of findings emerged (main effect for Informational Content, no main effect for Content Type, nor interaction in Experiments 1-3. An interaction did emerge in Experiment 4 for explicit attitudes, with follow-up tests indicating that negative attitudes were stronger in the Deepfake than authentic condition. A main effect also emerged for Content Type in Experiment 4 for IAT scores, such that automatic evaluations were larger in the Deepfake than authentic content condition (see footnote 10 on p.25).

**Reviewer 2**: Overall, I think the clarity and transparency of reporting data, analyses, and results could be improved significantly. The format for the methods section adopted here is unconventional and lacks clarity. For example, in Experiment 6, deepfake detection and awareness is measured, not manipulated. So, I was expecting some form of interaction result that shows participants' awareness x content type (genuine vs. deepfake) interaction on the DVs, and also results for positive and negative content. Furthermore, I wondered whether there were any systematic differences between the genuine vs. deepfake conditions in the percentage of participants who were aware of deepfakes. Similarly, I wanted to see the percentage of participants within each condition who accurately (vs. inaccurately) detected deepfakes (vs. genuine content). A chi-square that presents the percentage of participants who accurately (vs. inaccurately) detected deepfakes (vs. genuine) in each condition would be helpful.

**Authors**: We thank the Reviewer for pointing this point out to us. We have extensively revised the paper to better communicate our methods, analyses, and conclusions (see updated manuscript). We have also modified our OSF project page to make it easier for readers to navigate and interact with our various materials (<https://osf.io/f6ajb/>).

We have also added a new section outlining our analytic agenda and rationale in the confirmatory study (what was previously known as Experiment 6; see p.30-33). Finally, we have added a 2x2 confusion matrix depicting the number of participants who reported encountering genuine vs. Deepfaked content in the genuine vs. Deepfaked conditions (see Table 2 on p.35).

**Reviewer 2**: I wish the author team the best as they continue to develop this paper further. Thank you for the opportunity to read your paper.  
  
**Authors**: Many thanks for your kind words and constructive feedback. It really helped us when revising our manuscript and has hopefully led to a far stronger contribution.

**Reviewer 3**: The article "Deepfaked Online Content is Highly Effective in Manipulating Attitudes & Intentions" with a topic of growing societal relevance on how Deepfaked media influence people's attitudes and behavioral intentions. The experiments are well-designed, clearly built on each other, and provide relevant new insights. I was particularly impressed by the authors' efforts to produce their own deepfakes using one of the authors as the protagonist, which is a very clever way to sidestep ethical concerns of creating deepfakes of others. The writing is very accurate and to the point. Thanks to avoiding unnecessary repetitions in the main manuscript but providing very detailed descriptions of the experiments in the SOM, the paper is of ideal length. So overall, there is much to like about the paper.

**Authors**: We thank Reviewer 3 for their kind words, constructive comments, and for the Deepfaking papers they highlighted below. Their input helped us to produce what we believe is a stronger contribution.

**Reviewer 3**: At the same time, there are several concerns with the current version of the manuscript:

1. *Lack of relevant literature in intro and discussion*. After highlighting the societal relevance of deepfakes by referring to public media coverage, on page 5, the authors state, "What is needed then, alongside legislation and technological fixes, is a greater focus on the human dimension." I agree with that statement. At the same time, I was surprised to see the authors neglect several papers that have precisely done that.

Empirical work by Dobber et al. (2021) has studied the effect of micro-targeted fake news on political attitudes, and Vaccari & Chatwick (2020) test the deceptive potential of deepfakes. Also, when it comes to deepfake detection, research by Groh and colleagues has examined people's abilities to detect deepfake media for static images (Groh et al., 2021a), videos (Groh et al., 2021b); Other studies have tested different interventions to increase detection accuracy and uncovered cognitive biases in deepfake detection (Köbis et al., 2021a). Moreover, a special issue in Cyberpsychology, Behavior and Social Networking has been devoted to the social impact of deepfakes (see Hancock & Baileson, 2021). And conceptual work has highlighted the dangers of deepfakes (Köbis et al., 2021b).  
Including these relevant papers in the paper helps to embed the current study in the emerging stream of research on the social effects of deepfakes. Also, the statements in the discussion section about the novelty of the study (e.g., "Although politicians, journalists, academics, and think-tanks have all warned of the dangers that Deepfakes pose, our paper is one of the first to offer systematic empirical support for those claims.") need to be adjusted accordingly.

**Authors**: We sincerely thank the Reviewer for pointing out this literature to us. In line with this suggestion we have significantly revised the introduction to better highlight the on-going work taking place in psychology and computer science around Deepfaking (see changes on p.6-11). We have also adjusted the sentence in the General Discussion as requested.

**Reviewer 3**: *Measuring Implicit Attitudes (online)*. Combining measures of implicit and explicit attitudes is very useful. However, since the studies were conducted online, I wondered whether the IAT actually performs well enough. Of particular concern is whether the participants completed the study on a desktop or smartphone. Unfortunately, the paper does not mention previous research that has used IATs online nor does it provide detailed information about how such concerns of using the IAT online can be overcome.

**Authors**: One of the participation criteria for our studies was that participants completed the study on a laptop or desktop computer.

The IAT was also designed to be used online and the vast majority of IAT data collected over the past 20 years has been collected online via the Project Implicit website ([www.projectimplicit.net](http://www.projectimplicit.net)). A list of published studies that have used data from that website can be found here: <https://docs.google.com/document/d/1K1WnztJ2K3RPP5VOn6bDc0dr0ll1E3w0G2t6N4J3Dwo/edit>. Similarly, below the reviewer can find meta-analyses and systematic reviews of IAT data, including those collected online. In short, the measure was designed to be used in online settings and seems to work effectively in that context.

Greenwald, A. G., Poehlman, T. A., Uhlmann, E. L., & Banaji, M. R. (2009). Understanding and using the Implicit Association Test: III. Meta-analysis of predictive validity. *Journal of personality and social psychology, 97(1)*, 17.

Kurdi, B., Seitchik, A. E., Axt, J. R., Carroll, T. J., Karapetyan, A., Kaushik, N., ... & Banaji, M. R. (2019). Relationship between the Implicit Association Test and intergroup behavior: A meta-analysis. *American psychologist, 74(5)*, 569.

Greenwald, A. G., Brendl, M., Cai, H., Cvencek, D., Dovidio, J. F., Friese, M., ... & Wiers, R. W. (2021). Best research practices for using the Implicit Association Test. *Behavior research methods*, 1-20.

**Reviewer 3**: *Endogeneity concerns for detection and awareness measures*. One of the research questions the authors seek to answer is "does an awareness of Deepfaking and the ability to detect when it is present immunize them from its influence?" In the discussion section, they conclude that "A single brief exposure to a Deepfake quickly and effectively shifted attitudes and intentions, even when people were fully aware that content can be Deepfaked, and detect that they are being exposed to it." The way that detection and awareness are measured, however, undermine the validity of these conclusions.

When it comes to detection, the authors ask participants to admit whether they realized that the video was a deepfake after they told them it was a deepfake. This measure is problematic for several reasons. First, participants have no reason to answer this question honestly. In fact, they might misreport their responses, e.g., because they have image concerns such as appearing tech-savvy. Second, the open text format is unnecessary. If the measure is about deepfake detection, a binary Y/N answer would reduce noise in the measure. Third, letting authors who are not naïve to the hypotheses of the study code these answers increases this concern of noise further. Fourth, the detection measure is endogenous with the main outcome measures. These concerns similarly apply to the awareness measure. In the literature cited above, detection is typically assessed by showing deepfake and authentic videos and letting participants make (incentivized) guesses. Such a measure would better estimate the link between detection, awareness, and attitude change.

**Authors**: Reviewer 3 is correct. In two of our exploratory studies (2 and 4) informed participants what a Deepfake was, told everyone (even in the authentic condition) that they had just encountered one, and asked them if they were aware of this fact during the studies. We fully agree with the Reviewer that this way of probing detection is problematic for the various reasons they outline (e.g., subjective coding of open-ended data). With this in mind, we no longer unpack this ‘detection’ question in our revised manuscript.

Instead, we focus exclusively on the detection data from our confirmatory study (now called Experiment 5). We intentionally adopted a different (arguably) better approach in that study. Participants were first told the following:

“*Artificial Intelligence algorithms are now so advanced that they can fabricate audio and video content that appears real but was never said by a real person. This type of content is known as a ‘Deepfake’, and can be very convincing or difficult to tell from real content*.

*A key goal of this study is to examine whether people can tell the difference between genuine video content (footage of a real person) versus Deepfakes (videos created by computer algorithms that portray things that a person never said).*

*Some participants in this study were shown a genuine video of Chris. Other participants were shown a video of Chris where some sentences were Deepfaked (i.e., Chris never really said those things). It’s very important that you answer the following question honestly: Do you think that the video of Chris you watched earlier in this study was genuine or Deepfaked*?”

We then gave them two response options: “*The video I watched was Deepfaked: a computer algorithm was used to create footage of Chris saying things he never really said*.” And *“The video I watched was genuine: it only contained authentic video of an actual living person*.”

This approach circumvents many of the issues the Reviewer raised above. A binary Y/N approach was used that avoids the need for authors to subjectively code the data. Participants were *asked* what type of video they encountered rather than being told that they encountered one particular type. Inspection of the newly added 2x2 confusion matrix suggests that people did respond honestly. If they had “*misreport their responses, e.g., because they have image concerns such as appearing tech-savvy*” as the Reviewer suggests then we would have expected a very different distribution of responses (i.e., heavily skewed towards reporting Deepfakes in both the authentic and Deepfaked conditions which was not the case; see Table 2).

Nevertheless, we agree with the Reviewer that there are other ways of assessing and manipulating detection in the wider literature. We have included new material in the Introduction discussing work on Deepfake detection (see p.8-10) and General Discussion (p.47), highlight the limitations of the measure used in our confirmatory study, and propose future directions for research to address this (see p.51).

**Reviewer 3**: *Experimenter demand effects*. The set-up of watching a video and then rating the protagonist could appear artificial to some participants, especially since the statements could be viewed as unrealistic. Participants might therefore answer the attitude measures in the way they think the experimenters want them to. The authors appear to be aware of that concern and measure demand by asking participants in Experiments 1a-b about whether they replied in line with the experimenters' interests. I was left wondering what this measure actually revealed. Also, when it comes to "reactance" and "hypotheses guessing", I could not find information about the results of these measures, neither in the manuscript nor SOM. Adding information about these measures will help address readers' intuition that participants might have "played along" with the experimenters' demands.

As a side note, reporting at least the descriptive statistics of the other exploratory measures the authors took, e.g., over-claiming, would be beneficial too.

**Authors**: We see the Reviewer’s point here. Demand may have played a role in that participants were aware that they were in a psychology experiment and adjusted their responses on the outcome measure accordingly. This is actually one of the reasons why we included the IAT in our exploratory and confirmatory studies. This measure captures evaluative responses that are emitted automatically (in less than a second) and are thus less sensitive to demand than traditional self-report measures (for more on this see Gawronski, LeBel, & Peters, 2007; *Perspectives on Psychological Science*). We consistently and repeatedly found evidence for changes in attitudes at the implicit level, changes that mirrored those reported at the explicit level, which would argue against a simple demand explanation of our findings. That said, we have included additional material in the General Discussion highlighting ways that future work could better control for demand (e.g., by embedding Deepfaked content in even more naturalistic settings and measuring real world behavior in an unobserved manner; see changes on p.51).

With respect to the Reviewer’s question about the reactance, hypothesis guessing, and other exploratory measures. These were purely exploratory questions that all members of my lab are adding to all the experiments we run as part of a separate study on the role of these factors in attitude formation and change research. Unpacking and addressing these exploratory questions in the manuscript would require a sizeable investment in terms of time (to hand code many of the responses), and potentially detract from the main message we want to communicate in our paper. It is for this reason that we did not report the results of these questions in the manuscript.

However, in-line with the Reviewers request, we have now calculated basic descriptive data for many of the exploratory measures and individual difference questions (i.e., for those that did not require manual hand scoring). These can now be found in the “1. Exploratory\_Exp\_1-4\_Preregistered\_Analyses.html” file in our OSF project page. We direct interested readers to this file in the revised manuscript (see footnote 9 on p.20).

**Reviewer 3**: *Selection bias due to high exclusion rates*. In some of the studies, a relatively large proportion of participants were excluded, e.g., in Experiment 3, 55 out of 276 (~20%). Especially since these exclusions were not pre-registered, it can raise concerns about potential selection biases. An easy way to address this concern would be to report the results for the full sample and provide a more detailed rationale for the exclusion rule. This also applies to the non-pre-registered exclusions in Experiment 6.

**Authors**: We are slightly confused by this comment. Data were always excluded in-line with our pre-registered analytic plan. See the pre-registered “Data Analysis Plans” here:

<https://osf.io/z28nv> (i.e., Experiment 1 in the revised manuscript)

<https://osf.io/ka93u/> (i.e., Experiment 2 in the revised manuscript)

<https://osf.io/npb7g/> (i.e., Experiment 3 in the revised manuscript)

<https://osf.io/9xq4w> (i.e., Experiment 4 in the revised manuscript)

<https://osf.io/cjfrz> (i.e., Experiment 5 in the revised manuscript)

Nevertheless, we re-ran our analyses with the full sample (pre-exclusions) as requested by the Reviewer. The same set of findings emerged as reported post exclusions with one single exception: the *t*-test comparing the strength of attitudes established via Deepfaked video vs. audio became stronger (i.e., moved from *p* = 0.05 to *p* = 0.01). The Reviewer can inspect this for themselves in the OSF project page by navigating to the analyses folder and clicking on the “0. Exploratory\_Exp\_1-4\_Preregistered\_Analyses (Pre\_Exclusions).html” file.

**Reviewer 3**: *Lack of negative evaluation on IAT*. The IAT measures in the negative self-statements treatment indicate that participants do not (implicitly) evaluate the protagonist negatively. This is in contrast to the explicit measure where participants perceive the protagonist negatively in the negative self-statements treatments. This difference between the explicit and implicit attitudinal measures should be discussed.

**Authors**: We can understand the Reviewers point here. However, it is worth noting that the way the IAT is designed means that it produces *relativistic* scores rather than absolute ones. In other words, the zero point on the IAT does not reflect the absence of an attitude, or an ambivalent one. For a detailed treatment of the IAT’s zero point see the following paper:

Blanton, H., Jaccard, J., Strauts, E., Mitchell, G., & Tetlock, P. E. (2015). Toward a meaningful metric of implicit prejudice. *Journal of Applied Psychology, 100(5)*, 1468-1481.

It is for this reason that we interpret IAT scores in a relativistic manner in the manuscript (i.e., as more positive or more negative rather than positive or negative in some absolute sense).

**Reviewer 3**: *Lack of depth in the discussion section*. Although I appreciate the short length of the paper, the discussion section strikes me as too short and shallow. On top of neglecting the literature pointed out in comment #1, the discussion does also not address the limitations of the study. Besides the points raised above, discussion points could include the use of deception about the nature of the YouTube video (i.e., telling participants that Chris is trying to become a YouTube influencer) or the reliance on self-report measures (e.g., eliciting behavioral intentions instead of actual behavioral measures).

**Authors**: We have significantly revised and extended the General Discussion in line with the Reviewer’s suggestion. Specifically, we now add a Limitations section, point out many areas for future study, and situation our findings relative to the wider literature. See the General Discussion section for these changes.

**Reviewer 3**: Minor concerns:

* On page 10 of the paper, the author state, "Bob' had previously been used in our lab and shown to be evaluated neutrally during pilot testing." a reference to work backing this claim would be useful.

**Authors**: We don’t have a published paper to this effect. As such we have revised this sentence as follows: “A second individual (Bob) was selected from a large face database and served as the contrast category during the IAT.”

* Figure 2 is hard to grasp. Why not show the positive (upward) and negative (downward) effect for authentic (left) and deepfake (right) videos?

**Authors**: This figure has been removed from the revised manuscript.

* on page 22, more information about what Youden's J denotes would be useful  
  SOM:

**Authors**: We have revised the results section and provided more information on the classification statistics we used as requested (see changes on p.33, 34-35).

* add info about the length of the study and size of the remuneration to the SOM.

**Authors**: we have added information about the length of the study and rate of remuneration to the revised manuscript (see changes on p.15).

* on page 31, report SDs for the intention measures

**Authors**: SDs now reported.

* page 32 Lorah, 2018 in-text reference contains DOI that should be taken out

**Authors**: in-text reference removed.

* on page 33, add information about what MCMC sampling is

**Authors**: information and citation added

* page 34: "We used Gelman's (2019) method to characterize in order to characterize the priors as uninformative:" ◊ repetition of "to characterize"

**Authors**: repetition removed.