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**: Both the Editor and Reviewers raise questions about our work’s 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 also echo similar sentiments.

This argument 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 authentic content.

We agree with the logic behind this point. However the premise upon which the point is built is highly problematic. The vast majority of Deepfakes are not perfect replicas of authentic content but rather imperfect copies that vary drastically in their quality and believability. Most, including our own, contain numerous audio and visual artefacts (for several such examples, see <https://www.youtube.com/watch?v=ZJrffEfCMrs>).

These artefacts constitute perceptual cues which signal to the viewer that what they are watching or listening to has been tampered with, artificially constructed, or otherwise edited or modified. This should then lead the viewer to question the believability of what is being communicated to them and undermine the impact of that information on their thoughts, feelings, and actions.

In short, the vast majority of online Deepfakes are NOT perfect replicas of authentic content. Rather they are imperfect informational sources that contain cues which should trigger the perceptual system to question the validity of what one is viewing, reject that information, and its subsequent impact on ones attitudes and intentions.

When approached from this perspective, it is surprising that imperfect Deepfakes could quickly and powerfully shift (implicit) attitudes and intentions, and did so in comparable ways to authentic content. It’s also surprising that people who reported recognizing these artefacts (‘Deepfake detectors’) and who were aware of this technology both succumbed to its influence.

We have revised our manuscript to better highlight the above points (see changes on p.X).

**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 that first informed participants what a Deepfaked was, then 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. This style of phrasing may have influenced how participants responded. Second, the open-ended responses also introduced a degree of subjectivity into the scoring of these responses. Third, it may have encouraged 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 were addressing different issues).

We took each of these issues into mind and addressed them in our high-powered, confirmatory study by (a) asking rather than telling participants that they encountered a Deepfake, and (b) using a close ended Yes/No question. We continue to think that the exploratory detection analyses should be omitted from the revised manuscript. However, 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?*”. Note that participants did *not* know which condition they had been assigned to at any point in 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 video *may* have been Deepfaked after providing their evaluations.

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. It may be that factors related to the message (its emotional appeal) or the recipient (motivated reasoning) exert a stronger impact than those related to the content’s authenticity.

As we now acknowledge in the paper, this idea is consistent with previous research. For instance, many Instagram images are edited to create an idealized bodily image and viewing them can exert a negative psychological impact on the viewer. This negative impact persists even when disclaimers (cues) are attached to images explicitly calling into question their authenticity (e.g., Livingston, Holland, & Fardouly, 2020). Research on fake news also shows that people continue to believe and say they intend to share misinformation despite knowing it is being communicated by a low quality source (cue) (see Pennycook & Rand, 2021). Our findings indicate that Deepfakes also exert an impact on attitudes and intentions despite the presence of cues (visual and auditory artefacts) calling their authenticity into question.

We now discuss the above ideas and their contribution to the psychological literature throughout the revised manuscript (e.g., see footnote 2 on p.X and changes on p.X).

We have also integrated Reviewer 2’s suggestion for a design that would manipulate Deepfake detection awareness rather than simply measuring it (see changes on p.X) 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, 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*”.

Rather Deepfakes constitute one sub-type of a larger class of media known as *synthetic media* (i.e., content which is generated or manipulated via artificial intelligence techniques such as Generative Adversarial Networks [GANs] and related methods) (REF). Synthetic media can be divided into two broad categories. The first involves the use of AI to generate videos, images, audio, and text of individuals that do not exist (e.g., see <https://this-person-does-not-exist.com/en> or <https://www.resemble.ai/>) The second involves the use of AI to generate videos, images, audio, and text of individuals that *do* exist. It is this class of media that is commonly referred to as ‘Deepfakes’.

Reviewer 1 and 2’s points center on how one sub-type of synthetic media (Deepfakes) have often been *used* (e.g., to deceive viewers) and *properties* of that media (e.g., its believability or perceived authenticity). Note, however, that these very same criteria can be applied to any form of media (e.g., any form of video, text, audio, or photo can be used to deceive viewers and can vary in its believability or perceived authenticity). So these criteria do not define Deepfakes insofar as highlight how this form of synthetic media has typically been used.

As we note in our response to Reviewers 1 and 2, and as we have mentioned above, Deepfakes are often not “*manipulated so seamlessly*” to appear perfectly genuine (i.e., they can vary drastically in their quality and thus believability). Nor, as we now outline in the revised manuscript, are they always used to deceive viewers (e.g., Disney is using this type of media to create extras in its movies [REF], while industry is using it to generate fashion models [REF], or to protect the identities of persecuted minorities in online settings [REF]).

Finally, and as we noted above, the idea that Deepfakes are seamless replicas of genuine content that are designed to perfectly deceive viewers does not reflect the technology as it is used today or the vast majority of Deepfakes online.

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.

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.

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.

If you decide to revise the work, please submit a list of changes or a rebuttal against each point which is being raised (both here and below in the reviewers’ comments) when you submit the revised manuscript.  
   
What I expect is that you will take each of the concerns seriously and address these concerns in two ways. First, when possible, you should make changes in the manuscript to correct shortcomings that the reviewers perceive. (If there are comments that you do not find to be correct or apt, you still should consider that the incorrect perception is something that you might expect in other readers, so it would be helpful to take steps for the paper to anticipate such misperceptions and add clarifications in the text to prevent them.) The goal is to make your paper as accurate, scientifically responsible, interesting, and accessible to a wide range of experimental psychologists as possible.  
   
The second way that I would like you to respond to the reviewers is to put a lot of effort into a careful cover letter that goes through the comments point by point, explaining how you addressed each comment and, if you disagree with a comment, why you disagree (and, if possible, how you altered the writing in anticipation that other readers might have similar concerns).  
   
You might want to take advantage of a new policy as suggested in the editor’s  interview on the journal’s  web page: “I also encourage authors to include a brief paragraph at the end of the article describing its broader context, explaining such things as how the ideas originated, how the findings are related to the authors' research program, and how the research will be extended in the future, much as one provides in oral presentations.” You may label it something like Context of the Research if you wish. By brief, I mean about the same length as an abstract, not more than about 250 words.  
   
  
Sincerely,  
Adam Waytz  
Associate Editor  
Journal of Experimental Psychology: General  
 

**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!

**Authors**:

**Reviewer 1**: 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. In fact, 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**:

**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**: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**:

**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**:

**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**: *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**:  
  
**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**:  
  
**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**:  
  
**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**:  
  
**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**:  
  
**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**: 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**:  
  
I wish the author team the best as they continue to develop this paper further. Thank you for the opportunity to read your paper.

**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**:

**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**:

**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**:

**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.

**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**:

**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**:

**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**:

**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**:

**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**:

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

**Authors**:

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

**Authors**:

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

**Authors**:

* on page 31, report SDs for the intention measures

**Authors**:

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

**Authors**:

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

**Authors**:

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

**Authors**: