Journal of Experimental Psychology: General

Deepfaked Online Content is Highly Effective in Manipulating Attitudes & Intentions

--Manuscript Draft--

Manuscript Number:	XGE-2021-3737R1
Full Title:	Deepfaked Online Content is Highly Effective in Manipulating Attitudes & Intentions
Abstract:	Disinformation has spread rapidly through social media and news sites, biasing our (moral) judgements of individuals and groups. “Deepfakes”, a new type of AI-generated media, represent a powerful tool for spreading disinformation online. Although they may appear genuine, Deepfakes are hyper-realistic fabrications that enable one to digitally control another person’s appearance and actions. Across five studies (N = 2033) we examined the psychological impact of Deepfakes on viewers. Participants were exposed to either genuine or Deepfaked online content, after which their (implicit) attitudes and sharing intentions were measured. We found that Deepfakes quickly and effectively allow their creators to manipulate public perceptions of a target in both positive and negative directions. Many are unaware that Deepfaking is possible, find it difficult to detect when they are being exposed to it, and neither awareness nor detection serves to protect them from its influence. Preregistrations, data, and code are available at osf.io/f6ajb .
Article Type:	Unmasked Article
Keywords:	Deepfakes, AI-Generated Media, Implicit, Attitudes, Sharing Intentions, Detection
Corresponding Author:	Sean Hughes, Ph.D Ghent University Ghent, Henri Dunantlaan 2 BELGIUM
Corresponding Author E-Mail:	sean.hughes@ugent.be
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	Ghent University
Other Authors:	Ohad Fried
	Melissa Ferguson
	Ciaran Hughes
	Rian Hughes
	Xinwei Yao
	Ian Hussey
Author Comments:	
Corresponding Author's Secondary Institution:	
First Author:	Sean Hughes, Ph.D
Order of Authors Secondary Information:	
Manuscript Region of Origin:	BELGIUM
Suggested Reviewers:	
Opposed Reviewers:	
Order of Authors:	Sean Hughes, Ph.D
	Ohad Fried
	Melissa Ferguson
	Ciaran Hughes

	Rian Hughes
	Xinwei Yao
	Ian Hussey

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 1-3 for their kind words, as well as constructive and thorough feedback. It helped us when carrying out a major revision of our paper and has 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 breaks down as follows: 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 validity 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. Even imperfect Deepfakes 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 – as well as those used in past research - are NOT perfect replicas of authentic content; rather 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).

We also 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 5). 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).

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, we don't believe that its 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).

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. Importantly, not all Deepfakes are used for deceptive purposes (see p.5 for many relevant examples of Deepfakes being used for good).

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 does not 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 of how Deepfakes are being used for Good).

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 on 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 is exploring the use of Deepfakes for creating extras in its movies and tv series, 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 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). 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.

Cover Letter (only Editor will see)

Sean Hughes

E Sean.Hughes@UGent.be
T 0032 (0)9 264 86 49

Faculty of Psychology and Educational Sciences
Henri Dunantlaan 2
B-9000 Ghent
Belgium

DATE	PAGE	OUR REFERENCE
24/05/2022	1	Manuscript Resubmission

Dear Editor,

Please find enclosed our revised manuscript titled "Deepfaked Online Content is Highly Effective in Manipulating Attitudes and Intentions" considered for publication in *Journal of Experimental Psychology: General* (XGE-2021-3737_R1).

First and foremost, we would like to thank you as well as Reviewers 1-3 for your constructive feedback on our paper. We have taken time to carefully reply to each of your comments as well as those of the Reviewers. You can find these responses in an attached document.

All authors have approved the current version of the manuscript and made significant contributions to its writing and conceptualization. The manuscript meets the guidelines for ethical conduct and reporting of research, and holds no potential or actual conflicts of interest. It is not under review elsewhere; the data have not been published previously or accepted for publication. We thank you for your consideration and look forward to hearing from you soon.

Sincerely,

Sean Hughes
(*Corresponding Author*)

Ohad Fried
Melissa Ferguson
Ciaran Hughes
Rian Hughes
Xinwei Yao
Ian Hussey

Deepfaked Online Content is Highly Effective in Manipulating Attitudes & Intentions

Sean Hughes[1], Ohad Fried[2], Melissa Ferguson[3], Ciaran Hughes[4],

Rian Hughes[5], Xinwei Yao[6], & Ian Hussey[7]

[1] *Department of Experimental Clinical and Health Psychology, Ghent University, Belgium*

[2] *Interdisciplinary Center, Herzliya, Israel*

[3] *Department of Psychology, Yale University, USA*

[4] *Fermi National Accelerator Laboratory (Fermilab), USA*

[5] *Rudolf Peierls Centre for Theoretical Physics, Oxford University, UK*

[6] *Department of Computer Science, Stanford University, USA*

[7] *Faculty of Psychology, Ruhr University Bochum, Germany*

Author Note

Corresponding author: Sean Hughes, Department of Experimental Clinical and Health Psychology, Ghent University, Belgium. Email: sean.hughes@ugent.be. This research was conducted with the support of Grant BOF16/MET_V/002 to Jan De Houwer. Preregistrations, data,

and code are available at osf.io/f6ajb. SH conceptualized the studies, designed the methodologies, collected the data, contributed to data processing and analyses, wrote and reviewed the manuscript. OF and XY designed the Deepfaked videos. MF contributed to study conceptualization, reviewing and editing of the manuscript. CH and RH contributed to study conceptualization, data processing and analysis as well as reviewing and editing the manuscript. IH designed and implemented the data processing and analyses, contributed to study conceptualization, and reviewed the manuscript.

Abstract

Disinformation has spread rapidly through social media and news sites, biasing our (moral) judgements of individuals and groups. “Deepfakes”, a new type of AI-generated media, represent a powerful tool for spreading disinformation online. Although they may appear genuine, Deepfakes are hyper-realistic fabrications that enable one to digitally control another person’s appearance and actions. Across five studies (N = 2033) we examined the psychological impact of Deepfakes on viewers. Participants were exposed to either genuine or Deepfaked online content, after which their (implicit) attitudes and sharing intentions were measured. We found that Deepfakes quickly and effectively allow their creators to manipulate public perceptions of a target in both positive and negative directions. Many are unaware that Deepfaking is possible, find it difficult to detect when they are being exposed to it, and neither awareness nor detection serves to protect them from its influence. Preregistrations, data, and code are available at osf.io/f6ajb.

Keywords: Deepfakes, AI-Generated Media, Implicit, Attitudes, Intentions, Detection

Statement of Relevance

Conventional wisdom dictates that seeing is believing. However, thanks to recent advances in artificial intelligence, this may no longer be the case. A branch of machine learning known as 'deep learning' has made it increasingly easy to take a person's face, voice, or writing style, feed that data to a computer algorithm, and have it generate a synthetic copy. This 'Deepfake' can be used to convince others that what they are seeing, reading, or hearing is fact rather than fiction. Concern grows that Deepfakes pose a danger for the business, entertainment, intelligence, and political sectors. Across five studies we exposed people to Deepfaked online content and found that disinformation spread via this pathway is highly effective in manipulating the public's implicit and explicit attitudes as well as sharing intentions.

Deepfaked Online Content is Highly Effective in Manipulating Attitudes & Intentions

Oliver Taylor is a student at the University of Birmingham (UK) in his mid-twenties. He has brown eyes, dark hair, loves coffee and politics, and was raised in a traditional Jewish home. In his free time he serves as a freelancer and his work on anti-Semitism and Jewish affairs has been published in mainstream outlets such as the *Jerusalem Post* and *Times of Israel*. There's only one issue: Oliver doesn't exist. His university has no record of him, he has no digital fingerprint online, nor can anyone contact him. It appears that someone manufactured a false persona and coupled it with a hyper-realistic, AI-generated photo to propagate disinformation across international newspapers (Hill & White, 2020; Satter, 2020).

It turns out that Oliver is just the tip of a larger disinformation iceberg. Investigative reports have uncovered online networks of "political consultants" and "freelance journalists" propagating disinformation through conservative news outlets (Vincent, 2020). Facebook recently dismantled a network of fictitious "news editors from Kyiv" publishing pro-Russian propaganda (Gleicher & Agranovich, 2022), while others have uncovered networks of fake social media accounts of "regular people" spreading Chinese state propaganda (Burley, 2021) and disinformation favoring ex-Prime Minister Benjamin Netanyahu (Benzaquen, 2020). What unites these "political consultants", "freelance journalists", "news editors", and "genuine" social media users is that none of them exist.

Deception and disinformation are nothing new in mass communication. Text and image have long been manipulated for various ends: a politician or celebrity's comments can be edited and misreported while images on magazine covers, advertisements, and websites can be altered to depict their contents as better than they actually are. Although people are generally aware that text and images of others can be manipulated, watching others behave in a particular way on video, or hearing them state an opinion with their own voice, is still considered an accurate, valid, and trustworthy form of information.

However, this may no longer be true. A branch of artificial intelligence known as 'deep learning' has brought with it the ability to take another person's likeness (whether their face, voice, or writing style), feed that data to a computer algorithm, and have it generate a 'Deepfake': a hyper-realistic digital copy that can be manipulated into doing or saying anything. Many argue that Deepfakes represent a powerful new tool, one that in the wrong hands, can be used to supercharge deception as well as the spread of disinformation. [1]

Deepfakes are evolving at a rapid pace. Each year they become more realistic, easier to produce, and thanks to the Internet, distributed and shared on a mass scale (Kietzmann et al., 2020), with one report suggesting that they are doubling online every six months (Ajder et al., 2019). A variety of Deepfake apps freely exist online, allowing anyone with a computer or smartphone to swap one person's face with another (e.g., Zao, FakeApp, FaceSwap), manipulate what a target says (e.g., DeepFace Lab), take control of their voice (David, 2021), and even their writing style (Fagni et al., 2021).

Like any technology, Deepfakes can be used for both good and ill. Some are using it to generate believable voices and images for those who have lost their own through traumatic injury or cancer (Cattiau, 2019), and to digitally revive loved ones to help their family members deal with grief (Kim, 2022). Others are using the technology to enable celebrities such as David Beckham to deliver public health messages about malaria in multiple languages (Malaria Must Die, 2019), and museums to bring the dead back to life (e.g., visitors at the Salvador Dalí Museum can interact with a synthetic Dalí to learn about his art [Lee, 2019]).

Nevertheless, the technology also ripe for abuse. Deepfakes have quickly become a tool of harassment against activists (Satter, 2020), and a growing concern for those in the business,

[1] Deepfakes are a sub-category of AI-generated media that involve altering image, video, audio, or text to mimic and manipulate *existing* individuals (e.g., a well-known politician or a popular celebrity) (see Westerlund, 2019). Synthetic media is another sub-category of AI-generated media that involves generating and manipulating content of *non-existing* individuals. For instance, the technology can be used to generate images of people who do not exist (such as Oliver Taylor and the fake personas mentioned above; see Hill & White, 2020), synthetic voices that belong to no one (McDonough, 2020), and synthetic text that sounds human-authored (GPT3, 2020).

entertainment, and political sectors. The ability to control another person's voice or appearance opens companies to new levels of identity theft, impersonation, and financial harm (Bateman, 2020; Stupp, 2019). Female celebrities are being inserted into highly realistic pornographic scenes (Ajder et al., 2019), while worry grows that a well-executed video could have a politician 'confess' to bribery or sexual assault, disinformation that distorts democratic discourse and election outcomes (Galston, 2020; Koetsier, 2020). Elsewhere, intelligence services and think tanks warn that Deepfakes represent a growing cybersecurity threat, a tool that state-sponsored actors, political groups, and lone individuals could use to trigger social unrest, fuel diplomatic tensions, and undermine public safety (Sayler & Harris, 2020; Ciancaglini et al., 2020).

Recognizing these dangers, politicians in Europe and the USA have called for legislation to regulate a technology they believe will further erode the public's trust in media, and push ideologically opposed groups deeper into their own subjective realities (EU Commission, 2018; Cortez Masto, 2019). At the same time, industry leaders such as Facebook, Google, and Microsoft are developing algorithms to detect Deepfakes, excise them from their platforms, and prevent their spread (Burt & Horvitz, 2020; Canton Ferrer et al., 2020). Although legislative and technological stopgaps are undoubtedly necessary, they are also in a perpetual game of 'cat-and-mouse', with certain actors evolving new ways of evading detection and others rapidly working to catch up. In such a world, no law or algorithm can guarantee that the public will be completely protected from malicious synthetic content. What is needed then, alongside legislation and technological fixes, is a greater focus on the *human* dimension. It is imperative that we study how this new technology impacts our thoughts, feelings, and actions.

The Psychological Impact of Deepfakes

A small but growing literature indicates that Deepfakes may be able to manipulate attitudes, memories, trust in media, as well as intentions to share disinformation. In one study,

Hwang et al. (2021) found that disinformation communicated via Deepfaked videos was viewed as more vivid, persuasive, credable, and elicited stronger sharing intentions than disinformation which lacked such videos. Others have used Deepfakes to negatively manipulate attitudes of a known individual (politician) (Dobber et al., 2021) and to install false memories of fabricated events. Murphy and Flynn (2021) exposed participants to fake news stories via misleading text, misleading text and Deepfaked image, misleading text and Deepfaked video. A number of participants reported false memories following exposure to Deepfaked images or videos and that most viewed such content as convincing, dangerous, and unethical. [2]

Deepfakes may not only shift attitudes and memory but also trust in news and certainty about what is real. Vaccari and Chadwick (2020) exposed participants to a Deepfake of President Obama. Although viewers were not misled by that Deepfake they were more uncertain about what was real after watching it, and this uncertainty undermined their trust in news on social media. It may be that Deepfakes can be used to sow uncertainty about what is real and undermine collective belief in true events (a tactic used by state-sponsored propaganda campaigns; see Chadwick et al., 2018; Pomerantsev, 2015).

Preliminary work has also sought to determine who falls for and spreads Deepfaked content. One recent study examined if political interest, cognitive ability, and social network size influence Deepfake sharing online (Ahmed, 2021). Results indicated that *unintentional* sharing of Deepfakes was negatively associated with cognitive ability and positively associated with political interest, with the latter relationship moderated by social network size (i.e., the likelihood of sharing increased for politically interested citizens embedded in more extensive social networks). The same author also reported that *intentional* sharing of Deepfakes was

[2] It's worth noting that the Deepfake videos they selected (of celebrities and politicians) did not consistently increase false memory rates relative to misleading text or text-with-Deepfaked images.

positively associated with social media use and self-reported fear of missing out, and once again, was negatively associated with cognitive ability (Ahmed, 2022).

Finally, susceptibility to, and spread of, Deepfakes depends not only on the physical quality of the Deepfakes themselves but also congruence between pre-existing beliefs and communicated content. Shin and Lee (2022) examine how news articles containing a real or Deepfaked video influenced the credibility of, and intentions to share, that story. When there was a match between pre-existing attitudes and the content of the videos, people believed Deepfakes just as much as authentic videos and had higher intentions to share Deepfaked content (a finding consistent with work elsewhere on fake news; e.g., Sindermann et al., 2020).

Detecting and Mitigating Against Deepfakes

If Deepfakes do constitute a vivid, persuasive, and effective means of spreading disinformation, and are capable of shifting attitudes, memory, and trust in media, then we need to determine if (a) people can detect when they are exposed to such content, and (b) whether there are ways of minimizing its aversive effects. We briefly consider these topics below.

Detecting Deepfakes. The vast majority of work on Deepfake detection has centered on algorithms that can distinguish authentic from fabricated content (see Das et al., 2021; Groh et al., 2021; Katarya & Lal, 2020). The accuracy of AI detectors can vary from exceptionally high (> 95%) when benchmarked against their own training dataset under controlled settings to relatively low (~65%) when tested against the types of Deepfakes shared on social networks (i.e., those with varying compression levels, resizing, noise; Tolosana et al. 2020). A minority of this work has focused on whether humans can detect Deepfakes and how well they compare to their AI counterparts in doing so. Preliminary findings suggest that whether, and to what extent, humans detect Deepfakes varies as a function of the quality and type of Deepfake involved.

For instance, Rössler et al. (2018) exposed humans and AI to Deepfakes of differing quality (raw, low, high quality). AI methods were generally found to outperform humans, and human detection ranged from 69% accuracy for high quality Deepfaked images to 59% for low quality images (i.e., those images most likely to be shared on social media; also see Thaw et al., 2020). More recent work suggests that both human and AI detectors can be fooled by Deepfakes but that they tend to fall prey to different types (Korshunov & Marcel, 2021). This claim was corroborated by Groh et al. (2021) in a large-scale analysis of Deepfaked video detection by humans, AI, and hybrid human-AI collaborations. Analyses revealed that AI detectors beat individual humans but were similar to grouped human accuracy (suggestive of a 'collective intelligence' or 'wisdom of the crowds' effect). Whereas AI detectors were better able to detect Deepfakes involving novel unknown individuals, the opposite was true for known individuals (politicians), suggesting that humans and AI may be susceptible to different types of Deepfakes, with detection in humans likely informed by contextual information that extends beyond mere perceptional properties of the content. Interestingly, feeding humans an accurate prediction by the AI boosted their detection rates whereas feeding them an inaccurate prediction hindered those predictions (for related findings see Korshunov & Marcel, 2021). [3]

A similar story has emerged for Deepfake audio and text. For instance, AI detectors are generally superior to humans in detecting Deepfaked audio. Both humans and AI are tricked by different types of audio and that the most difficult Deepfake to detect for humans was easiest for AI and vice-versa (Muller et al., 2021). Audio detection in humans increased from 67% to 80% across early exposure and then plateaued despite with extended training. Younger participants were better able to detect fabricated audio than their older counterparts, while there was no correlation between IT-expertise and detection rates. Elsewhere, fake text reviews

[3] The growing realization that humans are more effective than AI when detecting certain forms of Deepfakes and that the latter are better than the former on others has led to the call for a hybrid human-machine detection system (e.g., Muller et al., 2021; although see Groh et al., 2021 for caveats on this).

generated by AI for products and services on platforms such as Yelp, TripAdvisor and Amazon successfully evade both algorithmic and human detection, and provide the same level of user-perceived "usefulness" as real reviews written by humans (Hovy, 2016; Yao et al. 2017).

Finally, several studies have examined if financial incentives, educational efforts, and individual difference factors boost Deepfake detection rates. In one study, participants were asked to distinguish Deepfaked from authentic videos, and motivated to do so via an educational intervention or financial incentive (Kobis et al., 2021). People were generally poor at differentiating Deepfaked from authentic content (~60% accuracy rates). Although the educational intervention and financial incentive increased motivation to detect it had no impact on participants actual ability to do so. Elsewhere, individuals scoring higher in social conservatism have been found to be more likely to endorse Deepfaked videos as authentic (Sütterlin, Ask, et al., 2022), while individuals with an IT or computer science background are no better than non-professionals at detecting Deepfakes (Sütterlin, Lugo, et al., 2022).

Mitigating Against Deepfakes. Several studies have sought to identify ways of mitigating against the negative impact of Deepfakes. Most have alerted participants (either before, during, or after exposure) to what Deepfaking is, or to the fact that they are coming into contact with one, and then measured the impact of the intervention on subsequent behavior. For instance, Iacobucci et al. (2021) primed participants with a definition of Deepfakes and their societal dangers before exposing them to one. Priming increased the rate at which Deepfaked content was successfully detected, with the relationship between priming and detection moderated by bullshit receptivity (i.e., a variable related to a less reflective cognitive style, lower cognitive ability, and tendency to believe in fake content; see Pennycook & Rand, 2020). Specifically, priming participants low in bullshit receptivity enhanced their ability to detect deceptive videos whereas this was not the case for their high scoring counterparts. Those who recognized a video had attempted to manipulate them were more likely to develop a

negative attitude toward that video, followed by a lower intention to share it with others. Elsewhere, Hwang et al. (2021) found that a general (as opposed to Deepfake specific) media literacy intervention reduced the perceived vividness, persuasiveness, and intention to share disinformation augmented by Deepfaked videos.

Others researchers have embedded disclaimers in Deepfaked content and found that content-with-disclaimers was less likely to be viewed as true and shared compared to content without disclaimers, especially for those higher in cognitive ability (Ahmed, 2021). Disclaimers also helped to minimize uncertainty and loss of trust in news on social media compared to Deepfaked content without disclaimers (Vaccari & Chadwick, 2020**)**. Finally, adding a post-exposure warning to help participants detect the fake stories did not stop a number of them from forming false memories based on that content (Murphy & Flynn, 2021). In contrast, post-exposure warnings have been found to reduce credibility and sharing intentions – but only when there is a match between pre-existing attitudes and the communicated content (Shin & Lee, 2022).

The Current Research

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, 2019). The rise of Deepfakes brings with it an increased risk of citizen impersonation and online abuse, especially towards the more 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). [4] It remains to be seen if a single brief exposure to Deepfaked online content is capable of manipulating implicit and explicit first impressions of members of the general public. 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.

To briefly preface what is to come, we first carried out four pre-registered exploratory studies (N = 1730) to examine how different types of Deepfakes (video vs. audio) created via different methods (cut-and-paste vs. fabricate-from-scratch) impacted implicit attitudes,

[4] Note that several studies have examined the public's attitudes *towards* Deepfaking as a technology (e.g., Murphy & Flynn, 2021). However this is significantly different from manipulating attitudes using Deepfakes.

explicit attitudes, and sharing intentions. Thereafter, a high-powered confirmatory study (N = 635), replicated our initial work while testing the three questions outlined above. Taken together, this work provides the first systematic empirical examination of how Deepfakes shape our implicit and explicit first impression of others.

Experiments 1-4: Exploratory Studies

Experiments 1-4 set out to test a widely held yet empirically unverified assumption: that Deepfaked content can bias both implicit and explicit attitudes. In Experiments 1-2 we focused on Deepfake *videos*. Participants navigated to YouTube where they watched a clip of a novel individual (Chris) disclose personal information about himself. Half watched him emit three positive and two neutral self-statements while the other half watched him emit three negative and two neutral self-statements. Self-reported and automatic attitudes were then assessed. By manipulating the informational *Content* participants encountered we sought to demonstrate that Deepfakes can be used to shift public perceptions in both positive and negative directions.

Our second and key manipulation centered on the *Type* of content participants came into contact with, such that half were exposed to an authentic recording of Chris emitting the aforementioned statements while the other half watched a Deepfake of him. In Experiment 1 Deepfakes were generated using a 'cut-and-paste' approach wherein the target's words and actions were extracted ('cut') from one video and then inserted ('pasted') into another video (see Fried et al., 2019). In Experiment 2 Deepfakes were fabricated from scratch to simulate situations where authentic content is not available or cannot be obtained and has to be generated whole cloth (Yao et al., 2020). In either case the Deepfake had Chris 'confess' to either virtuous (positive attitude induction) or malicious actions (negative attitude induction). [5]

[5] Both techniques are ripe for abuse. For instance, the 'cut-and-paste' method can be used to extract statements made by a political candidate from one video (e.g., where they genuinely talk about the dangers that climate change pose), modify them, and insert them into a completely different video, where they now appear to warn about the dangers that racial outgroups pose. It can also be used to 'scrape' publicly available content from a person's social media account and Deepfake them for malicious (blackmailing) purposes. The 'fabricate from

Manipulating the *Type* of content participants viewed allowed us to address two related questions: (a) can Deepfakes alter our implicit and explicit attitudes, and if so, (b) are they similar, better, or worse than authentic content in doing so? Note that if one begins from the position that Deepfakes are perfect replicas of authentic content then these questions may seem self-evident. If one cannot distinguish fabricated from authentic content then surely the former will be treated as equivalent to the latter, and both types of media should impact attitudes and intentions in similar ways. Yet we caution against such a position. Although there has been an exponential increase in the quality of Deepfake content over the years, they still vary drastically in their respective quality and believability. Most, including our own, contain clearly detectable audio and visual artefacts (e.g., Shin & Lee, 2022). These artefacts should signal that the content one is viewing or listening to has been tampered with, artificially constructed, or otherwise edited. This may undermine the believability of the information and its subsequent impact on one's attitudes and intentions. Thus its worth asking *if* Deepfakes can shift attitudes and intentions *despite* the presence of cues highlighting that the content may be suspect, and if so, how attitudes induced in this way compare to those established via authentic content. [6]

Experiments 3-4 extended our analyses from one type of Deepfaked media (video) to another (audio). A similar procedure was used as before but with one important difference: video clips were now substituted for audio clips. Authentic audio clips were created by extracting audio from the authentic videos used in Experiment 2. Deepfake audio clips were generated by feeding a training set of the target's voice to a bidirectional text-to-speech

scratch' method can be used to place words in the mouth of a political opponent, discredit activists, or for purposes of identity theft or impersonation (e.g., Bateman, 2020).

[6] Research indicates that media can exert a psychological influence over the recipient even when cues are present that undermine the validity and thus believability of what is being communicated. For instance, many social media 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 et al., 2020). Research on fake news shows that people can continue to believe and intend to share misinformation despite knowing it is being communicated by a low quality source (see Pennycook & Rand, 2021). Thus it may be that Deepfakes also exert an impact on attitudes and intentions despite the presence of cues (artefacts) calling their authenticity into question.

autoregressive neural network (Mason, 2019). This process allowed the neural network to learn how to mimic the target's voice. The end result was a Deepfaked voice that sounded similar to the target and which could be manipulated by the researcher into saying anything. By cloning the target's voice and manipulating what he 'said' we sought to determine if Deepfake audio would manipulate attitudes and intentions towards the target, and whether it does so in comparable ways to Deepfake video. [7]

Method

Sample Size Selection

Samples were selected on a convenience basis for Experiments 1-4.

Participants and Design

The demographic breakdown for the exploratory studies can be found in Table 1. Participants took part via the Prolific website (https://prolific.ac) in exchange for a monetary reward at a rate of £5 per hour with a typical completion time of 15mins. Assignment to the informational *Content* (positive or negative attitude induction) and Content *Type* conditions (Authentic vs. Deepfake content) was counterbalanced across participants and both served as independent variables. An additional method factor was also counterbalanced across participants (self-reported ratings vs. IAT first) but was not included in our analytic models. Ratings and IAT scores were the dependent variables of interest. Study designs and data-analysis plans for all experiments are available on the Open Science Framework website (osf.io/f6ajb). We report all manipulations and measures used in our experiments. All data were collected without intermittent data analysis. The data analytic plan, stimuli, materials, data, and

[7] If Deepfaked audio can manipulate our attitudes and intentions then this would provide a cheaper, less resource intensive, and more widely available method of psychological manipulation than video content. The fact that hackers recently Deepfaked a CEO's voice and used it to trick an employee into initiating a six-figure wire transfer supports the idea that Deepfaked audio may have such an effect (Stupp, 2019; also see Tran, 2021).

deviations from pre-registration are available at the above link. [8] Ethical approval was granted for all studies by the Ethical Commission within the Faculty of Psychology and Pedogeological Science at Ghent University, Belgium (2020/135).

Table 1. Demographic information for Experiments 1-4.

Experiment	Sample Size	Gender	Age	
		(*Female*)	**M**	**SD**
1	428	232	30.7	9.0
2	276	151	32.6	12.3
3	429	258	30	8.6
4	265	154	33.3	12.6

Stimuli

Attitude Objects

An unknown individual (Chris) served as the target during the attitude formation phase (this individual was the first author who was selected on the basis of convenience). The target appeared during the video or audio while his images also served as one set of category stimuli during the pIAT. A second individual (Bob) was selected from a large face database and served as the contrast category during the IAT.

Behavioral Statements

Eight self-statements were selected for use in the videos and audio: three positive, three negative, and two neutral. These items were selected from a larger pool that were pre-tested along three dimensions: valence, believability, and diagnosticity.

[8] We carried out two additional exploratory studies that are not reported here. These studies were precursors to Experiments 1-4 and sought to determine if authentic videos are capable of changing implicit and explicit attitudes towards a target individual. We found that authentic videos indeed led to strong changes in implicit and explicit attitudes (see osf.io/f6ajb).

Personalized IAT (pIAT)

A set of eight positive and eight negative trait adjectives were used as valenced stimuli during the pIAT in Experiments 1-4. The names of two individuals (Chris and Bob) served as target labels and the words 'I like' and 'I dislike' as attribute labels. Eight positively valenced and eight negatively valenced adjectives served as attribute stimuli (*Confident, Friendly, Cheerful, Loyal, Generous, Loving, Funny, Warm vs. Liar, Cruel, Evil, Ignorant, Manipulative, Rude, Selfish, Disloyal*) while images of the two individuals served as the target stimuli.

Authentic Content

Authentic positive videos consisted of three positive self-statements and two neutral statements, while negative videos consisted of three negative and two neutral self-statements. Authentic audio clips were created by extracting the audio content from the authentic videos used in Experiment 2.

Deepfaked Content

In Experiment 1, Deepfakes were created using a 'cut-and-paste' method. Broadly speaking, this involves 'cutting' a target's genuine content from one video and 'pasting' it into another video. In our case we took the authentic videos, fitted a parameterized 3D model to the target's head, and then used this model to generate computer graphical renderings of his face and mouth movements. These renderings were then converted to photorealistic synthesized video using a trained Generative Adversarial Network and served as the raw input for the Deepfakes (for more on this method see Fried et al., 2019). Specifically, a Deepfaked negative video was created by removing the positive statements from the authentic positive video and inserting in the Deepfaked negative statements, while a Deepfaked positive video was created by removing the negative statements from the authentic negative video and pasting in the Deepfaked positive statements. These fabricated videos were then uploaded to YouTube where participants watched them.

In Experiment 2 we used the "fabricate from scratch" method developed by Yao et al. (2020). Instead of using only 3D model parameters from existing data of the actor, this method leverages both a small amount of the actor's data as well as a large repository of speaking footage of a different actor to generate high quality 3D head model parameters for arbitrary spoken content. It also allows easy iterative editing. Given recordings of only the negative statements, we used this method to iteratively perform localized edits (i.e. word or short phrase replacements) on clips of negative statements until they were transformed into their positive counterparts. At each iteration, we spliced in real audio recordings of the actor to obtain the audio for that iteration. Deepfaked videos of the actor saying negative statements were generated similarly. In this way videos were similar in their content but differed in their origin (see Figure 1).

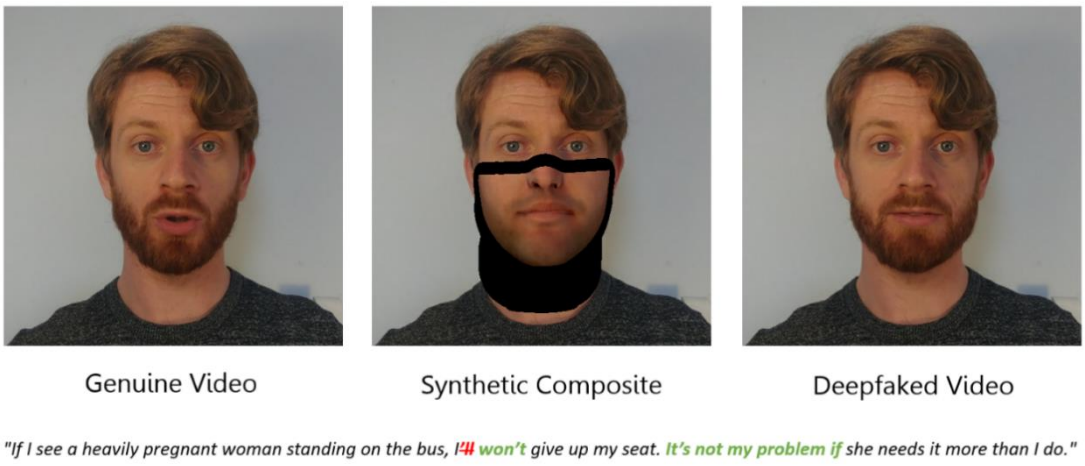

Figure 1. Deepfake creation method used in Experiment 2 ('*fabricate from scratch*'). This approach leverages a small amount of the target's genuine data as well as a large repository of speaking footage of a different individual to generate high quality 3D head model parameters for the desired Deepfaked content. This approach allowed us to transform genuine positive statements into Deepfaked negative statements and genuine negative statements into Deepfaked positive statements, thereby controlling how the target was perceived and how others intended to interact with him.

Procedure

Participants were welcomed to the study and asked for their informed consent. Studies generally consisted of four sections: demographics, attitude induction phase, attitude assessment phase, and exploratory questions.

Demographics

Age and gender information was obtained in Experiments 1-4, along with country of residence, ethnicity, educational level, employment status, and income in Experiments 2 & 4.

Attitude Formation Phase

Participants were first told the following: "In this study we are interested in how people remember and react to what they see online. You are going to watch a video (listen to audio: Experiments 3-4) taken from a YouTube channel. The person who makes these videos (audio) is called Chris. Please watch Chris' video (listen to Chris' audio) and pay close attention to what he says. We will ask you questions about this later on."

Thereafter participants navigated to YouTube where they watched a video (or listened to an audio recording). During the video (audio), the target emitted three valenced self-statements as well as two neutral statements. Half of the participants encountered positive variant video/audio wherein he emitted three positive and two neutral statements, whereas the other half encountered a negative variant video/audio, wherein he emitted three negative and two neutral statements. For half of the participants the content they encountered was authentic, while for the other it was Deepfaked (see osf.io/f6ajb/ for video and audio clips used).

Attitude Assessment Phase

Implicit Attitudes

Following the attitude induction phase, a personalized IAT (pIAT) was administered to measure implicit attitudes towards the target (Chris) relative to an unknown individual (Bob). Participants were informed that they would encounter two individuals (Chris and Bob) as well as the words 'I like' and 'I dislike' (attributes) which would appear on the upper left and right sides of the screen, and that stimuli could be assigned to these categories using either the left ('F') or right keys ('J'). If the participant categorized the image or word correctly the stimulus disappeared from the screen and, following a 400ms inter-trial interval (ITI) the next trial

began. In contrast, an incorrect response resulted in the presentation of a red 'X' which remained on-screen for 200ms, and was followed by an ITI and the next trial (for a detailed overview of the pIAT's block structure see Supplementary Materials).

Self-Report Attitudes

Self-reported ratings of Chris were assessed using three Likert scales. On each trial, participants were presented with a picture of Chris and asked to indicate whether they considered him to be 'Good/Bad', 'Positive/Negative' and whether 'I Like Him/I Don't Like Him along a scale that ranged from -3 to +3 with 0 as a neutral point.

Sharing Intentions

In Experiment 4 participants were asked to indicate how they intended to behave with respect to the target ("1. If I were browsing YouTube and encountered Chris' video, I would support him by clicking the 'share' button [i.e., share his video with other people]"; "2. Chris has just started to make these videos and wants to become a YouTuber. I happen to encounter his video on YouTube. I would 'subscribe' to his channel to learn more about him." "3. I would recommend Chris' videos to others"). Responses were emitted using a scale ranging from -2 (*Strongly Disagree*) to 2 (*Strongly Agree*) with 0 (Neutral) as a center point.

Individual Difference Measures

A number of individual difference measures were taken in Experiment 2, including measures of political ideology, religiosity, cognitive ability (revised cognitive reflection test [rCRT]), preference for effortful or intuitive thinking styles (rational-experiential inventory [REI]), overclaiming, conspiratorial thinking, deepfake awareness and detection. Preference for effortful vs. intuitive thinking (REI), and cognitive ability (rCRT) were also taken in Experiment 4. The over-claiming and conspiratorial thinking measures were replaced in Experiment 4 with a news evaluation task (i.e., a measure of people's ability to discern real from fake news; familiarity with those news stories and their willingness to share them) as well

as a measure of actively open-minded thinking (Actively Open Minded Thinking – Evidence) (for additional information on each of these measures see Supplementary Materials). [9]

Exploratory Questions

Questions related to content memory, diagnosticity, demand, reactance, hypothesis, and influence awareness were included for exploratory purposes. These questions were not central to the research agenda in our exploratory studies and are not discussed from this point onwards. We have made this data freely available at (osf.io/f6ajb) for those interested in examining it further.

Results

Participant Exclusions

We screened-out participants who (a) failed to complete the entire experimental session and thus provided incomplete data and/or (b) who had IAT error rates above 30% across the entire task, above 40% for any one of the four critical blocks, or who complete more than 10% of trials faster than 400ms (n = 70 [Experiment 1], n = 55 [Experiment 2], n = 88 [Experiment 3], n = 47 [Experiment 4]). This led to a final sample of 358 in Experiment 1, 221 in Experiment 2, 341 in Experiment 3, and 218 in Experiment 4.

Data Preparation

Self-report ratings from the three Likert scales were collapsed into a mean score with positive values indicating positive attitudes towards Chris and negative values the opposite. Response latency data from the IAT were prepared using the D2 algorithm recommended by Greenwald et al. (2003). Scores were calculated so that positive values reflected a relative implicit preference for Chris whereas negative values indicated the opposite. We also

[9] It quickly became apparent that questions about the relationship between demographic, individual difference factors, attitudes, and deepfake detection was itself a separate line of work, and one that extended beyond the remit of our research agenda. As such, these additional measures were not analyzed in this paper (but still reported for transparency purposes). We have also made all data and analyses related to demographic and individual difference factors available to others (see osf.io/f6ajb).

calculated an evaluative change score in order to examine if the videos led to a change in evaluations regardless of *Information Content* (positive vs. negative statements). We did so by reverse scoring self-reported ratings and pIAT scores for those in the negative video conditions. Positive values indicated a change in attitudes in the predicted direction, negative values indicated the opposite, whereas neutral values indicated an absence of an attitude or ambivalence.

Analytic Strategy

A series of *t*-tests were carried out on the rating and IAT data (dependent variables) to determine if that data differed as a function of Informational *Content* (positive vs. negative behavioral statements) (i.e., we examined for a main effect of Informational Content). A series of independent and one-sample *t*-tests were also carried out on the ratings and pIAT data to determine if they differed as a function of *Content Type* (authentic vs. Deepfaked) (i.e., we examined for a main effect of *Content Type*). Cohen's *d* are reported for all of the comparisons. Bayes factors in accordance with procedures outlined by Rouder, Speckman, Sun, Morey, and Iverson (2009) were also examined in order to estimate the amount of evidence for the hypothesis that there is a difference in evaluations as a function of Informational *Content* or *Content Type* (alternative hypothesis) or no such difference (null hypothesis).

Hypothesis Testing

Deepfake Videos are Highly Effective in Manipulating Implicit and Explicit Attitudes

Experiment 1: 'Cut and Paste' Method. Results revealed that explicit attitudes towards Chris differed as a function of informational *Content*. Participants exposed to positive information reported positive attitudes towards the target (M = 1.35, SD = 1.27) whereas those exposed to negative information reported negative attitudes (M = -1.69, SD = 1.47), $t(318.43) = 20.62$, $p < .001$, $d = 2.22$, 95% CI [1.96; 2.49], $BF_{10} > 10^5$. This was also the case for implicit attitudes which also varied as a function of exposure to positive (M = 0.39, SD = 0.31) vs.

negative content (M = 0.04, SD = 0.36), t(317.27) = 9.92, p < .001, d = 1.07, 95% CI [0.85; 1.29], $BF_{10} > 10^5$.

Critically, for our purposes, Deepfakes created via the 'cut and paste' method were highly effective in manipulating how the target was perceived, both at the explicit: M = 1.51, SD = 1.38, t(176) = 14.58, p < .001, d = 1.09, 95% CI [0.91; 1.28], $BF_{10} > 10^5$; and implicit levels: M = 0.19, SD = 0.41, t(176) = 6.11, p < .001, d = 0.46, 95% CI [0.31, 0.61], $BF_{10} < 10^4$). Finally, analyses revealed that the self-reported attitudes, t(355.83) = -0.10, p = .92, d = 0.01, 95% CI [-0.22; 0.20], BF_{10} = 0.12, and implicit attitudes, t(353) = 0.52, p = .60, d = 0.06, 95% CI [-0.15; 0.26], BF_{10} = 0.13, induced by Deepfakes were similar in magnitude to those induced by authentic content (i.e., attitudes did not differ as a function of *Content Type*).

Experiment 2: 'Fabricate from Scratch' Method. Attitudes once again differed as a function of Informational *Content*. Participants exposed to positive information reported positive attitudes (M = 1.36, SD = 1.27) while those exposed to negative information reported negative attitudes towards the target (M = - 1.65, SD = 1.34), t(212.9) = 17.12, p < .001, d = 2.31, 95% CI [1.97; 2.66], $BF_{10} > 10^5$. This was also the case for implicit attitudes which also varied as a function of exposure to positive (M = 0.40, SD = 0.29) vs. negative content (M = 0.03, SD = 0.31), t(212.04) = 9.34, p < .001, d = 1.26, 95% CI [0.97; 1.55], $BF_{10} > 10^5$.

Deepfakes fabricated from scratch were highly effective in manipulating how the target was perceived, both in terms of people's explicit attitudes (M = 1.41, SD = 1.31, t(108) = 11.22, p < .001, d = 1.08, 95% CI [0.84; 1.31], $BF_{10} > 10^5$) and implicit attitudes (M = 0.23, SD = 0.34, t(108) = 6.84, p < .001, d = 0.65, 95% CI [0.47, 0.84], $BF^{10} > 10^4$). Deepfakes also led to similar strength attitudes as authentic content, both at the explicit, t(218.79) = -1.01, p = .32, d = -0.14, 95% CI [-0.39; 0.13], BF_{10} = 0.24, and implicit levels, t(216.69) = 0.95, p = .35, d = 0.13, 95% CI [-0.14; 0.39], BF_{10} = 0.22 (i.e., attitudes did not differ as a function of *Content*

Type). Experiment 2 therefore replicated our initial findings and demonstrated that they hold for different Deepfake creation methods.

Deepfaked Audio is Highly Effective in Manipulating Public Perceptions of a Target

Experiments 3 & 4. A similar pattern of findings emerged for Deepfake audio as observed for Deepfake video. Explicit attitudes varied as a function of Informational *Content*, such that Chris was liked more after listening to the positive compared to negative audio clips, both in Experiment 3 (positive: M = 1.35, SD = 1.05 vs. negative: M = -1.86, SD = 1.23), $t(330.86) = 25.92$, $p < .001$, d = 2.81, 95% CI [2.51; 3.11], $BF_{10} > 10^5$, and Experiment 4 (positive: M = 1.51, SD = 1.01 vs. negative: M = -1.85, SD = 1.31), $t(186.84) = 20.91$, $p < .001$, d = 2.87, 95% CI [2.47; 3.26], $BF_{10} > 10^5$. The same was true at the implicit level, with Chris automatically preferred following positive compared to negative audio clips, both in Experiment 3 (positive: M = 0.40, SD = 0.28 vs. negative: M = 0.05, SD = 0.31), $t(335.69) = 11.18$, $p < .001$, d = 1.21, 95% CI [0.98; 1.44], $BF_{10} > 10^5$, and Experiment 4 (positive: M = 0.39, SD = 0.31 vs. negative: M = -0.06, SD = 0.35), $t(200.89) = 9.93$, $p < .001$, d = 1.36, 95% CI [1.06; 1.66], $BF_{10} > 10^5$. Sharing intentions towards Chris (Experiment 4) were ambivalent following positive information (M = -0.39, SD = 0.96) and highly unfavorable after negative information (M = -1.58, SD = 0.74), $t(213.23) = -10.32$, $p < .0001$, d = -1.38, 95% CI [-1.67, -1.08], $BF_{10} = > 10^4$.

Critically, analyses revealed that Deepfake audio was highly effective in manipulating public perceptions of the target at the explicit (Experiment 3: M = 1.54, SD = 1.24, $t(172) = 16.26$, $p < .001$, d = 1.24, 95% CI [1.04; 1.43], $BF_{10} > 10^5$; Experiment 4: M = 1.89, SD = 1.06, $t(111) = 18.82$, $p < .001$, d = 1.78, 95% CI [1.48; 2.08], $BF_{10} > 10^5$) and implicit levels (Experiment 3: M = 0.17, SD = 0.36, $t(172) = 6.22$, $p < .001$, d = 0.47, 95% CI [0.32, 0.63], $BF_{10} > 10^4$; Experiment 4: M = 0.23, SD = 0.38, $t(111) = 6.84$, $p < .0001$, d = 0.61, 95% CI

[0.41, 0.81], $BF_{10} > 10^4$). This was also true when sharing intentions were measured in Experiment 4, $t(111) = 4.78$, $p < .0001$, $d = 0.45$, 95% CI [0.27, 0.64], $BF_{10} > 10^4$.

Finally, Deepfakes led to self-reported attitudes of similar magnitude as authentic audio in Experiment 3, $t(335.41) = 1.09$, $p = .28$, $d = 0.12$, 95% CI [-0.10; 0.33], $BF_{10} = 0.21$, and even larger attitudes than authentic audio in Experiment 4, $t(206.7) = 2.92$, $p = .004$, $d = 0.39$, 95% CI [0.13; 0.67], $BF_{10} = 7.95$. There was no difference in the magnitude of implicit attitudes (Experiment 3: $t(337.26) = -0.37$, $p = .71$, $d = -0.04$, 95% CI [-0.25; 0.17], $BF_{10} = 0.13$; Experiment 4: $t(216) = -0.18$, $p = .85$, $d = -0.03$, 95% CI [-0.29; 0.24], $BF_{10} = 0.15$), or sharing intentions as a function of *Content Type* (Experiment 4: $t(215.04) = 0.75$, $p = .45$, $d = 0.10$, 95% CI [-0.16; 0.37], $BF_{10} = 0.19$). [10]

Attitudes Induced via Deepfaked Videos Are Similar to Those Induced via Deepfaked Audio

We combined the data for participants exposed to Deepfakes in Experiments 1-2 (video) as well as those in Experiments 3-4 (audio). We then compared if the magnitude of attitude induction varied as a function of *Media Type* (video vs. audio). Analyses revealed that Deepfake videos led to similar changes in explicit, $t(560.16) = 1.93$, $p = 0.05$, $d = 0.16$, 95% CI[-0.003; 0.33], $BF_{10} = 0.57$, and implicit attitudes, $t(568.11) = 0.22$, $p = 0.82$, $d = 0.19$, 95% CI[-0.18; 0.15], $BF_{10} = 0.10$, as Deepfaked audio.

Interim Discussion

The findings from Experiments 1-4 revealed that Deepfakes can quickly and powerfully impact viewers, equipping their creators with a means of controlling how others are perceived

[10] In our pre-registered analytic plan we stated that we would carry out a series of *t*-tests examining for a difference in attitudes as a function of Informational *Content* or *Content Type*. However, during the review process, a reviewer asked if we would consider the main and interaction effects between the above variables. With this in mind, we conducted a 2 (*Content*: positive vs. negative) x 2 (*Content Type*: authentic vs. Deepfake) ANOVA on the ratings and IAT scores. Analyses from Experiments 1-3 mirrored those reported above: a main effect emerged for Informational *Content*, but no main effect for *Content Type* nor an interaction between the two. In Experiment 4 the interaction term was significant for explicit attitudes, $F(1, 214) = 9.97$, $p = .002$, $\eta_p^2 = 0.05$, 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.

at the implicit and explicit levels. This is true despite their audio and visual artefacts, and for different types of Deepfaked content (video and audio), different Deepfake creation methods ('cut and paste' vs. 'fabricate from scratch'), and different psychological outcomes (explicit attitudes, implicit attitudes, and sharing intentions).

Experiment 5: Confirmatory Study

We carried out a high-powered, pre-registered, confirmation study to provide an even stronger test of the hypotheses from Experiments 1-4. Specifically, we set out to confirm our two core hypotheses: that even imperfect Deepfakes can quickly and powerfully shift attitudes and intentions towards a target (H1) and that they are as effective in doing so as authentic content (H2). We additionally had several new hypotheses that were either induced from, or refined based on, our exploratory studies and which required testing.

For instance, we wanted to know if people can detect when they have been exposed to Deepfaked content (H3). If they were aware of the concept of Deepfaking prior to the study, and if this awareness increased their chances of detecting a Deepfake when it is present (H4). Similarly, we were curious if prior awareness of Deepfaking (H5) or correctly detecting its presence (H6) would help 'immunize' people from its influence, and if those who were both aware *and* who reported detecting the Deepfake were better immunized than those who are not (H7). To answer these new questions we replicated Experiment 2 (Deepfaked videos) while making improvements to the design, preregistration specificity, and analytic strategy (e.g., we swapped to a Bayesian framework to produce more intuitive effect sizes and tests of non-inferiority). Experiment 5 therefore set out to confirm that Deepfakes can be used to manipulate (implicit) attitudes and intentions, and to explore if people are aware of this new technology, can detect when they are being exposed to it, and if awareness and/or detection helps protect them from its influence.

Method

Sample Size Selection

Sample size was determined via Bayesian power analysis which was itself determined using a simulation study. The simulation involved the following steps. Bayesian linear models were first fitted to the data from our exploratory studies to provide point estimates of the parameters used in these hypothesis tests. These parameters were then used to simulate data that met the same 'true' parameters. The models were then refit to the simulated data, and hypothesis tests were applied. 1000 iterations of this "simulate-data-fit model-test hypotheses" process were then performed. The proportion of simulations which detected the known 'true' effects (i.e., statistical power) was then summarized. The number of participants simulated was varied between simulation runs until a sample size was obtained that provided at least 80% power for all hypotheses. This sample size was then adjusted to take the data exclusion rates observed in our exploratory studies into account. Results indicated that 600 participants would be required after exclusions.

Participants and Design

770 participants completed the study on Prolific in exchange for a monetary reward. Data processing was run on this sample to determine if the following criteria were met: at least 600 participants remaining after exclusions (for H1 and H2), at least 166 participants who were shown a Deepfake and reported prior awareness of Deepfaking (for H5), at least 103 participants who were shown a Deepfake and correctly detected it as a Deepfake (for H6), and at least 46 participants who were shown a Deepfake, reported prior awareness of Deepfaking, and correctly detected it as a Deepfake (for H7). These sample size requirements were derived from the power analysis via simulation study to provide power > .80 for each hypothesis.

The final (post-exclusion) sample consisted of 635 participants (387 female, M_{age} = 35.7, SD = 13). Informational *Content* (positive vs. negative) and *Content Type* (authentic vs.

Deepfake) were counterbalanced between participants, and were used as Independent Variables in the analyses. Evaluative task order (ratings vs. pIAT first) was also counterbalanced but not modelled in analyses.

Stimuli

A similar set of stimuli were used as in Experiment 2.

Procedure

Participants completed the tasks in the stated order unless it was previously noted that a given phase was counterbalanced.

Demographics

Participants indicated their age and gender (man, woman, non-binary, prefer not to disclose, prefer to self-describe).

Acquisition Phase

Participants watched the same authentic or Deepfaked videos as in Experiment 2.

Personalized IAT

A similar pIAT was used as before with one exception: pIAT trials were increased from 16 to 20 in the practice blocks and 32 to 40 in the test blocks.

Self-Reported Ratings and Intentions

A similar set of rating and intention questions were used as in previous studies. Intention ratings now ranged from -3 (*Strongly Disagree*) to +3 (*Strongly Agree*).

Deepfake Detection

At the end of the study participants were 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?"

Participants were given two closed-ended response options: "The video I watched was Deepfaked: a computer algorithm was used to create footage of Chris saying things he never really said" or "The video I watched was genuine: it only contained authentic video of an actual living person". They were also asked to "Please give a reason for your answer in the text box below", and provided with a means to indicate their open-ended response. This open-ended question was included for exploratory purposes and not used in our preregistered analyses.

Deepfake Awareness

Prior awareness of Deepfaking as a concept was assessed using the following question: "Prior to this study did you know that videos could be 'Deepfaked'? Two closed-ended response options were provided (Yes - I was aware of the concept of Deepfakes / No - I wasn't aware of the concept of Deepfakes). Participants were asked to "Please elaborate on your answer using the text box below" and provided with an open-ended response option. This open-ended question was included for exploratory purposes and not used in preregistered analyses.

Results

Data Exclusions

Data were excluded as in Experiments 1-4 with one addition: we now excluded participants if they spent too little (< 2.25 minutes) or too much time (> 4.5 minutes) on YouTube. These exclusion lengths were selected on an analysis of video linger times from our

exploratory studies and selected to exclude individuals who failed to spend sufficient (or who spent excessive) time on the video (n = 68).

Data Preparation

Data were prepared as before. This time we also standardized self-reported ratings, pIAT scores, and sharing intentions by 1 SD after exclusions and prior to analyses. This was done within each level of both IVs (i.e., as a function of informational *Content* [positive vs. negative], or *Content Type* [authentic vs. Deepfaked]).

Analytic Strategy

As we noted above, we implemented a Bayesian framework to better formalize our core research questions, hypotheses, analytic models, inference rules, and other researcher degrees of freedom. This analytic strategy is described below and was designed to provide strong tests of our various hypotheses. We specify how each of our verbal hypotheses correspond to a statistical inference rule that would be used to conclude support for that hypothesis.

All evaluative dependent variables (self-reported ratings, IAT D2 scores, and sharing intentions) were standardized (by 1 SD) after exclusions and prior to analysis condition (see Lorah, 2018). This was done within each level of both IV (i.e., by information *Content* [positive vs. negative], and *Content Type* [authentic vs. Deepfaked]). As such, the beta estimates obtained from the Bayesian linear models represent standardized beta values. The nature of this standardization makes these estimates somewhat comparable to the frequentist standardized effect size metric Cohen's *d*, as both are a difference in (estimated) means as a proportion of SD - although they should not be treated as equivalent. Effect size magnitude here can be thought of as using comparable scales as Cohen's *d*. As such, to aid interpretability, the point estimates of these beta estimates are reported as δ (delta) rather than β.

Models

Bayesian Models. Bayesian models were implemented using the R package brms (Buerkner, 2017), which leverages the STAN language to allow for Bayesian inference via MCMC sampling.

Linear Models. The linear models (hypotheses 1, 2, 5, 6, 7) took the following generic format: a dependent variable (IAT score, ratings, or sharing intentions); two independent variables, information *Content* (positive vs. negative) and *Content Type* (authentic vs. Deepfaked); and their interaction. Wilkinson notation: dependent_variable ~ information_content * content_type.

Poisson Model. The Poisson model (hypothesis 4) took the following format: cell counts served as dependent variable; two independent variables (Deepfake concept awareness and Deepfake detection); and their interaction. Wilkinson notation: counts ~ awareness * detection.

Model Priors and Their Informativeness

Wide priors have been specified for all parameters (i.e., normal distribution with $M = 0$ and $SD = 10$, following general recommendations for weakly informative priors in STAN: https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations). We used Gelman's (2019) method to characterize the priors as uninformative: For each parameter, we compared the posterior SD to the prior SD. If the posterior SD for any parameter was more than 0.1 times the prior SD, we noted that the prior was informative, otherwise it was noted as uninformative. Inspection of prior and posterior distributions for the models fit to the data from our previous experiments (1-4) allowed us to conclude that all priors were uninformative. As such, results (i.e., derived from posterior distributions) were very weakly influenced by the prior, and therefore likely to be comparable to what would be found had we used frequentist estimation methods (i.e., driven in large part by the data rather than the prior).

Model Convergence

We inspected the convergence of the chains via visual inspection of the plots, $\hat{R}$, and the effective sample size metrics. Appropriate changes to model hyper parameters were made if evidence of non-convergence was found (e.g., increasing number of iterations or the adapt_delta parameter and refitting the model).

Parameter Estimation and Inference

Posterior distributions were summarized via a metric of central tendency, the Highest Maximum A Posteriori probability estimate (MAP). This was judged to be a preferable metric to the mean given the mean's sensitivity to outliers. Estimation width was quantified via 95% Credible Intervals via asymmetric Highest Density Intervals (HDIs). In the linear models, estimates for subgroups were calculated via manipulation of the posterior probabilities (e.g., authentic condition = intercept, Deepfaked condition = intercept + main effect for experiment condition, etc.; see R code implementation for details).

Bayesian p values were also produced for the sake of familiarity for many readers. These were derived from the proportion of the posterior samples that were in the predicted direction: Bayesian $p = (1 - P(Beta > null))/2 \approx$ frequentist p value (where $null$ refers to $\delta = 0$ in the linear models or IRR = 1 in the Poisson model). All three of these metrics were implemented using the bayestestR R package.

Null-Hypothesis Test

Null-hypothesis tests (e.g., for H1, H4, and H5) were implemented via the inspection of the 95% Credible Intervals. If a CI's lower bound was > $null$ (where $null$ refers to $\delta = 0$ in the linear models or IRR = 1 in the Poisson model), this was considered evidence in support of the alternative hypothesis (e.g., that the estimated means differed).

Non-Inferiority Tests

Non-inferiority tests (e.g., for H2) were implemented via the general method described by Lakens, Scheel, and Isager (2018), albeit (1) applied to intervals derived from Bayesian models and (2) applied unidirectionally (i.e., as a non-inferiority rather than equivalence test). Specifically, if the lower bound of the 95% CI of the authentic video condition was < the lower bound of the 90% CI of the Deepfaked video condition (i.e., the difference between information *Content* conditions in each subgroup), this was considered evidence in support of the alternative hypothesis (i.e., evidence of non-inferiority in estimated means; that Deepfakes are as good as authentic content).

In addition to this non-inferiority hypothesis test, which we note is a relatively strict test, an effect size was produced to characterize the magnitude of the effect size in the Deepfaked condition as a percentage of the authentic condition. This was implemented by calculating a proportion for each posterior sample and then parameterizing this new distribution (via MAP and 95% HDI). In addition to the above non-inferiority test, we concluded that Deepfaked content produces substantively similar attitude effects (in a continuous rather than categorical sense) by describing this estimate of comparative effect size (e.g., that the magnitude of the Deepfake condition was within ± 10% of authentic content).

Classification Statistics

Many have argued that no single classification metric is optimal. Therefore a confusion matrix and multiple classification metrics were calculated using the true status of the video content (authentic or Deepfaked) and participants Deepfake detection responses, specifically: False Positive Rate, False Negative Rate, Balanced Accuracy, and Informedness (Youden's *J*). 95% Confidence Intervals were bootstrapped using the case removal and percentile methods and 2000 iterations.

Hypothesis Testing

Deepfakes Can Be Used to Manipulate (Implicit) Attitudes and Sharing Intentions

Analyses confirmed that self-reported attitudes (δ = 2.38, 95% CI [2.15, 2.58], p < .0000001), implicit attitudes (δ = 1.37, 95% CI [1.14, 1.57], p < .0000001), and sharing intentions (δ = 1.69, 95% CI [1.49, 1.92], p < .0000001) all differed depending on the informational *Content* encountered in the Deepfake videos. Put another way, Deepfaked content shifted attitudes and intentions in a positive direction when positive attitude induction was applied and in a negative direction when negative attitude induction was applied. [11]

Deepfakes Influence Attitudes and Intentions to a Similar Extent as Authentic Content

Self-reported attitudes induced by Deepfaked content were slightly inferior to those established via authentic content (genuine lower 95% CI = 2.37; Deepfake lower 90% CI = 2.15). Stated more precisely, Deepfaked videos were 91.1% (95% CI [80.2, 103.1]) as effective in changing self-reported attitudes compared to authentic content. A different pattern emerged for implicit attitudes (pIAT scores): Deepfaked videos were non-inferior to authentic content (genuine lower 95% CI = 1.16; Deepfake lower 90% CI = 1.14), and was 96.9% (95% CI [75.9, 120.9]) as effective in changing implicit attitudes as authentic videos. Finally, behavioural intentions induced by Deepfaked content (lower 90% CI = 1.49) were non-inferior to those established via authentic content (lower 95% CI = 1.50), with Deepfakes 97.9% (95% CI [80.5, 116.8]) as effective in changing intentions as authentic videos.

People Find It Difficult to Detect Deepfakes

At the end of the study we outlined what a Deepfake was and asked participants if the YouTube video they had just watched was genuine or Deepfaked. Our aim here was to

[11] A similar main effect of informational *Content* emerged for authentic videos as well: self-reported attitudes (Standardized effect size δ = 2.60, 95% CI [2.37, 2.81], p < .0000001), implicit attitudes (δ = 1.38, 95% CI [1.16, 1.61], p < .0000001), and sharing intentions (δ = 1.72, 95% CI [1.50, 1.95], p < .0000001).

determine if people can successfully detect what type of content they had actually been exposed to. To answer this question we first computed a 2 x 2 confusion matrix to determine the *True Positive* (Authentic content classified as authentic), *False Positive* (Authentic content classified as a Deepfake), *True Negative* (Deepfaked content classified as a Deepfake), and *False Negative* rates (Deepfaked content classified as authentic) (see Table 2). We then used this information to compute *Sensitivity* or how many authentic videos were correctly classified as authentic (i.e., True Positives / True Positives + False Negatives = 0.63) and *Specificity* or how many Deepfaked videos were correctly classified as Deepfaked (i.e., True Negatives / True Negatives + False Positives = 0.65).

Table 2. Confusion matrix of self-reported judgements about the nature of the content participants encountered during the study.

		Judgement Type	
Content Type		**Reported Genuine**	**Reported Deepfake**
	Actually Genuine	*True Positive (TP)* 184	*False Positive (FP)* 120
	Actually Deepfake	*False Negative (FN)* 109	*True Negative (TN)* 221

We then computed two classification statistics. The first was the Balanced Accuracy statistic (i.e., Sensitivity + Specificity / 2) which is useful for determining how accurate classification is when there are potential imbalances in the four fields of the confusion matrix. Analyses revealed that participants did not make *accurate* decisions about the type of content they had encountered (Balanced Accuracy = 0.64, 95% CI [0.60, 0.67]), far lower than what might be considered as a highly accurate decision (BA of .80 or .90).

The second classification statistic was Youden's *J* which indicates the likelihood that a person will make an "informed decision" as opposed to a random guess. Youden's *J* combines Sensitivity and Specificity into a single measure ([Sensitivity + Specificity] – 1) and can range from 0 to 1 (i.e., a value of 1 indicates that there are no false positives or false negatives [perfect detection of *Content Type*] whereas a value of 0 indicates the same proportion of positive results in both conditions [a totally useless test]). Analyses revealed that participants were poor at making *informed* decisions about Content Type (Youden's $J = 0.27$, 95% CI [0.20, 0.35]), far less than what might be considered a highly informed decision (*J* of .80 or .90).

People Who Are Aware That Content Can Be Deepfaked Are Better Able to Detecting Them

We also wanted to know if prior awareness that Deepfaking was possible would serve to protect viewers when they were eventually exposed to a Deepfake themselves. To answer this question we first asked participants if they were aware of Deepfaking as a technology before taking part in our study. 56% of our sample reported such an awareness. We then selected those participants who had encountered a Deepfake video in our study and computed an incidence rate ratio (IRR). This statistic allowed us to compare the incident rate (i.e., how likely a Deepfake was correctly classified as a Deepfake) between two groups: (a) those who reported prior awareness that Deepfaking was possible, and (b) those who reported no such awareness. We also used a Bayesian Poisson model to estimate a 95% Credible Interval around the effect's Incidence Rate Ratio. Analyses revealed that people who were aware of the concept of Deepfaking before participating in our study were 1.9 times more likely to detect they had been exposed to a Deepfake than those who encountered a Deepfake and lacked this prior awareness (IRR = 1.92, 95% CI [1.45, 2.51], $p < .001$). Specifically, those who were previously unaware of Deepfaking had a 23% chance of detecting they had been exposed to one whereas their aware counterparts had a 44% chance of detection.

Prior Awareness Does Not Protect One From Being Influenced by Deepfakes

We then examined if attitudes and sharing intentions would still emerge for 'aware' participants (i.e., those who were exposed to a Deepfake and who reported being aware of the concept of Deepfaking prior to taking part). Results indicated that prior awareness of Deepfaking as a technology did not protect an individual from being influenced by a Deepfake. The self-reported attitudes, δ = 2.10, 95% CI [1.83, 2.41], $p < .0000001$, implicit attitudes, δ = 1.32, 95% CI [1.03, 1.59], $p < .0000001$, and sharing intentions, δ = 1.50, 95% CI [1.22, 1.81], $p < .0000001$, of these individuals were altered in-line with the Deepfake content.

Deepfake Detection Does Not Protect People From Being Influenced

The aforementioned analyses centered around people who reporting being aware of Deepfaking *prior* to the study. We were also curious to know if 'Deepfake detectors' (i.e., those who correctly recognized that they had encountered a Deepfake *during* the study) would be more immune to the influence of that content (i.e., show smaller changes in attitudes and intentions). To examine this question we selected the 'Deepfake detectors' from our sample and examined how they responded. Results showed that self-reported attitudes, δ = 2.19, 95% CI [1.93, 2.44], $p < .0000001$, implicit attitudes, δ = 1.38, 95% CI [1.11, 1.62], $p < .0000001$, and sharing intentions, δ = 1.59, 95% CI [1.33, 1.84], $p < .0000001$, all varied in accordance with Deepfaked content.

The Combined Impact of Awareness & Detection Does Not Protect Against Influence

Finally, we wanted to know if viewers who were both aware of Deepfakes prior to the study *and* who successfully detected the presence of the Deepfake, would be immune to their influence. Results indicated that this was not the case: Deepfake detectors who were also aware of the technology prior to the study still showed expected changes in self-reported attitudes, δ = 1.98, 95% CI [1.65, 2.28], $p < .0000001$, implicit attitudes, δ = 1.35, 95% CI [1.01, 1.66], $p < .0000001$, and sharing intentions, δ = 1.40, 95% CI [1.08, 1.72], $p < .0000001$.

Interim Discussion

A high-powered, pre-registered, confirmatory study replicated the core findings from our exploratory studies: Deepfakes can be used to manipulate (implicit) attitudes and intentions, and did so in similar ways to authentic content, despite their imperfections. Many participants are unaware of this new technology, find it difficult to detect when they are being exposed to it, and neither awareness nor detection served to protect them from its influence.

General Discussion

Politicians, academics, and think-tanks all warn that Deepfakes represent a dangerous new tool for the creation and spread of disinformation, one with implications for democratic discourse, national and private-sector security, as well as citizens attitudes, memories, and trust in media (e.g., Hwang, 2020; Van Huijstee et al., 2021). Whereas attention has largely focused on legislative and technological solutions to the dangers of Deepfaking, less has been said about the human dimension. A small but growing literature has emerged on the *Psychology of Deepfakes*, and illustrated how AI-generated media has the potential to influence our thoughts, feelings, and actions; that people often find it difficult to differentiate fact (authentic content) from fiction (Deepfakes); and that several factors may help people to detect and mitigate against this new technology.

Our work contributes to this burgeoning literature in three ways. First, it shows that a single brief exposure to a Deepfake can influence implicit attitudes, explicit attitudes, and sharing intentions towards regular members of the public. The total control afford by the technology allowed us to bias the recipient's perceptions of the target, so that he was strongly despised by some and favored by others. This was true for both audio and video Deepfakes. Somewhat surprisingly, Deepfake-induced attitudes and intentions were similar in strength to those established via authentic content. This was despite the fact that our Deepfakes – like most used in past work and currently circulating online - contained audio and visual artefacts that

should have alerted recipients to the fact that this content had been manipulated, and thus constitutes a flawed or biased information source. Our findings indicate that even imperfect Deepfakes can exert an immediate and powerful impact on the public's perceptions. We will return to this point below.

Second, we show that people find it difficult to detect when they are being exposed to Deepfakes of members of the general public. Classification statistics revealed that participants failed to make accurate or informed decisions about the type of content they had encountered. Nearly half of our confirmatory sample (44%) were unaware that Deepfaking was possible prior to the study while accurate Deepfake detection rates were also far from perfect (65%). These rates are similar to those reported elsewhere in the literature (e.g., Groh et al., 2021; Kobis et al., 2021; Rössler et al., 2018; Thaw et al., 2020), and suggest that people can struggle to recognize imperfect Deepfakes. [12] Finally, we show that prior awareness and/or detection does not immunize people from being biased by Deepfaked content. Participants whom we thought would have the greatest potential to reject Deepfakes as biased informational sources (i.e., those who were previously aware of Deepfaking and successfully detected that they had been exposed to one) still had their attitudes and intentions shaped in the targeted directions.

In what follows we consider the implications of these findings as well as open questions and future directions for research in this area.

Theoretical Considerations

Perhaps the most interesting question is why Deepfakes influenced (implicit) attitudes and intentions despite containing visual and auditory artefacts. These artefacts represent validity cues signaling that the content had been tampered with, artificially constructed, or

[12] Note that we (a) assessed Deepfake detection using a single post-hoc question at the end of the study, and (b) never informed participants about the type of content they would encounter during the study. This strategy differs from how detection has been studied in other studies (e.g., asking participants to determine whether a series of images, videos, or audio clips are authentic or Deepfaked).

otherwise edited or modified. This should have undermined the believability of that content and its subsequent impact on the recipient.

We can see several possibilities here. The first is that participants were aware of the artefacts within the Deepfaked content but failed to stop and reflect on the implications of those cues or integrate this validity information into their evaluative judgements and decision making (sharing intentions). Popular dual-process theories posit that human cognition can be broadly divided into two qualitatively distinct categories with opposing properties: 'intuitive' cognitive processes that are emitted quickly, without intention, awareness, or control (System 1), and 'deliberate' cognitive processes that are emitted slowly, with intention, awareness, and control (see Evans & Stanovich, 2013; De Neys, 2021). [13] In situations where motivation and opportunity to think in deliberate and effortful ways is low, people may over-rely on 'quick-and-dirty' thinking (System 1) that eschews careful deliberation in exchange for speed and efficiency. Others have argued that humans are 'cognitive misers' who seek to minimize resource intensive cognitive efforts, and as a result, often fail to consider all aspects of the content they encounter, instead focusing only on certain salient features (i.e., System 1 thinking may be a default setting for many people; Fiske & Taylor, 2013).

If correct then people may focus more on salient aspects of Deepfaked content such as the target's appearance and demeanor or the information being conveying and less on those subtle validity cues which require careful and analytic consideration. Put simply, it's not that people *cannot* think analytically about Deepfaked content, it's just that they opt not to do so. Such over-reliance on System 1 thinking may be exacerbated under automaticity conditions that deprive the individual of the opportunity or motivation to stop and critically evaluate what they are seeing and hearing. These conditions are likely present in the contexts where

[13] Despite its popularity the System 1 vs. 2 dichotomy has increasingly been challenged as an over simplification of human cognition and the conditions under which it occurs (for more see De Houwer, 2019; De Neys, 2021).

Deepfaked content is encountered (i.e., when one is quickly scrolling through their social media feed or news website in the presence of multiple distracting stimuli, and without the intention to carefully evaluate each piece of content encountered).

The above perspective is consistent with past work showing that people often fall for others forms of disinformation (fake news) because they fail to think rather than lack the ability to do so (Pennycook & Rand, 2019). Fake news is also less likely to be believed by people who are more reflective and when people are actively asked to stop and deliberate on the accuracy of that content (Bago, Rand, & Pennycook, 2020). If this also holds for Deepfakes, then analytic thinking may play a "corrective" role in helping recipients to override their automatic, intuitive responses to fabricated content, and to detect as well as integrate validity cues into their evaluative judgements and decision making. Future work could examine if people with analytic thinking styles or abilities are more skeptical and less impacted by Deepfakes, or if interventions that train those abilities, or have people slow down and engage in reflective reasoning, reduces the endorsement and spread of disinformation transmission via Deepfakes.

Another possibility is that people may not only fail to critically evaluate Deepfaked content but also over-rely on simple mental heuristics that increase the chances that visual artefacts are discounted or not even recognized. For instance, visual information inspires greater trust and confidence (e.g., Newman et al., 2015; Sherwin et al., 2006), is processed more easily (Stenberg, 2006) and leads to greater cognitive elaboration than text-based content (Lazard & Atkinson, 2015; Seo et al. 2020). Video and images also tend to spread on Twitter more than text-based messages (Goel et al., 2015, p.186). Such findings support the idea of a 'realism heuristic' wherein audio-visual content which bears a close resemblance to the real world is processed more easily and trusted to a greater extent than text based content (i.e., "seeing is believing"; Sundar, 2008).

Elsewhere, research on the truthiness effect shows that adding thematically related visual information (image, video) to text can systematically bias people into believing content to be true (e.g., Newman & Zhang, 2020). It may be that audio-visual content boosts the ease and speed with which the recipient can extract meaning and comprehend communicated content (i.e., processing fluency) and that this metacognitive experience increases perceived truth relative to other types of media (text). If so, then audio-visual Deepfakes may exploit this heuristic and increase the chances that people discount artefacts and thus fall for disinformation more than they would for text-based content (e.g., "these visual or audio artefacts are probably due to some technical issue with the video rather than clues that the content itself has been fabricated"). This may help explain why the audio and video Deepfakes both exerted a strong impact on attitudes and intention. It would also suggest that audio-visual Deepfakes may be a particularly persuasive form of communication, especially when compared to text-based (disinformation) pathways such as fake news propagated through websites and Twitter (c.f. Murphy & Flynn, 2021). Future work could examine this possibility by comparing transmission pathways in terms of their relative effectiveness.

Another salient feature of the Deepfakes we used was their emotional evocativeness. The target either confessed to virtuous (e.g., "*Most of my weekends are spent helping my grandmother around her house. She is really old and I want to spend as much time with her as possible before she passes on*") or repugnant acts (e.g., "*I won't give up my seat on the bus if I see a heavily pregnant woman standing. It's not my problem if she needs it more than I do*") that were designed to provoke positive reactions or shock, anger, and moral outrage. Emotion may have functioned as a heuristic that directed attention away from validity cues in the Deepfake (artefacts) and towards the information being communicated, thus undermining the former's integration into judgements and decision making. Something similar has been found

in the context of fake news: emotionally evoking content can increase belief in, and spread of, misinformation (e.g., Martel et al., 2020; Vosoughi et al., 2018).

A final possibility is that participants were asked to evaluate a novel target they had no prior knowledge of, and were only given a single piece of information to go on (the Deepfaked content). In the absence of any other relevant information they may have defaulted to what they knew despite the presence of cues that elicited concern about the quality of that information. In other words, they may have applied the Gricean maxim of relation (i.e., "I am being told this information for a reason: it must be relevant to the task at hand"; Grice, 1975), and used the only information they had to form an attitude towards the target. This explanation may apply to Deepfakes of novel targets to greater extent than those of known individuals (where prior knowledge about the target maybe used to inform one's judgements). If this explanation holds, it would suggest that people applied the maxim despite the presence of validity cues questioning the quality of what was being communicated. In other words, rather than recognize the communicated content was flawed and declining to make a judgement until more accurate information was available, they instead relied on the only information they had at hand to formulate an evaluative judgement.

Open Questions and Future Directions

Much has been written about Deepfaking and its societal implications in the popular media. Yet many of the claims being made have never been empirically verified while others are based on a single proof-of-concept paper that requires replication and extension. It's possible that Deepfakes may re-write our memories (Resnick, 2018), fool our brains (Smith, 2019), undermine our collective trust in media (Fortson, 2022), or democratic discourse (De Witte et al., 2022). But these and many other claims require systematic and rigorous research before they can be trusted. Research that takes the *sender* of the Deepfake, the *content* being

communicated, the *recipient* of the Deepfake, and the wider *context* into account. We see many open questions and directions for future work in this area.

Attitudes. The systematic investigation of Deepfakes and their ability to establish new, and revise existing, attitudes is still in its infancy. We still don't know if Deepfakes can change explicit and implicit attitudes towards known individuals such as celebrities, politicians, or industry leaders, in both positive and negative directions. [14] The extent to which these attitudes are predictive of subsequent changes in real-world behavior also remains to be seen. For instance, do Deepfake-induced attitudes towards politicians influence the public's support for policies connected to that individual, their voting intentions, or the politician's chances of being (re)elected? If a Deepfaked celebrity communicates the "risks" associated with vaccines or the virtues of a homeopathic medicine will this impact health behaviors such as vaccine uptake or product purchases? The same goes the general public: would the online circulation of a negative Deepfake influence one's job hiring or online dating prospects, social standing, membership to a political party or other group? Would a Deepfake of a suspect committing a crime influence real or mock jurors decisions (Grothaus, 2021)? We recommend that future research assess for both attitude and behavioral change when establishing and revising Deepfake-induced attitudes.

We also recommend that future work consider the longevity and durability of these attitudes. Researchers could replicate our work and assess for attitude change immediately after induction, and then again after a brief (48 hours) or long-term (month) delay to determine if Deepfakes have a persistent or short-lived impact. Others could examine if attitudes induced via authentic content can be altered via Deepfakes or vice-versa. For instance, a positive attitude could be induced towards a novel individual using a Deepfake, participants then

[14] Only one paper that has empirically examined if Deepfakes can be used to change attitudes towards a known individual so far (Dobber et al., 2021). The authors of that paper conducted a single exploratory study, focused solely on self-report measures, induced attitudes in just one (negative) direction, towards a Dutch politician.

informed that the content they had just encountered was fabricated, and thereafter be exposed to authentic negative content to see if the original attitude can be updated. Alternatively, participants could be shown authentic negative content of a known individual followed by Deepfaked positive content to see if the latter can reverse the reputational damage caused by the former. Manipulating the order (induction vs. revision) and type (authentic vs. Deepfaked) of content people come into contact would provide valuable insight into the malleability of attitudes to disinformation. Researchers could draw on recent insights from the field of attitude updating when carrying out this work (e.g., the finding that updating is maximally effective when content is highly diagnostic, believable, and leads the recipient to reinterpret previously encountered content; see Ferguson et al., 2019). [15]

Memory. It also remains to be seen if Deepfakes can manipulate what we remember at the individual or collective level, either by installing false memories of events that never happened (i.e., Mandela effects) or by altering genuine memories that did (Liv & Greenbaum, 2020). Preliminary evidence suggests that Deepfakes may have the potential to install false memories but systematic study is needed to strengthen this claim, especially research indicating when, why, and how they have this effect (Murphy & Flynn, 2021). We are aware of no study examining if Deepfakes can be used to revise existing memories. This could be achieved by exposing participants to footage of historical figures discussing events that never occurred (e.g., President Nixon discussing the failed moon landing; DelVisco, 2020) or to Deepfaked movies that people have previously watched (e.g., would people endorse Keanu Reeves as having acted in Forrest Gump after watching a Deepfake of him doing so). Future research could also determine if Deepfakes can be used to influence memory during the creation,

[15] Setting questions about deception and disinformation to the side, Deepfakes also present attitude researchers a with a flexible new tool that can use to create stimuli which are perfectly matched and tailored to their experimental needs (e.g., the same target individual can be manipulated into emitting any statement, thus providing sophisticated control over the content they want to investigate). This could be used by those studying any aspect of attitude formation and change wherein the attitude object is a known or novel person, or a message communicated by those individuals.

storage, and retrieval phases. Or if a false memory created at one moment in time, and then triggered at a later moment in time, influences the recipient's future behavior. For instance, imagine that a damaging Deepfake of a known politician is widely circulated in the months prior to an election. Could the memory of this be triggered by a political opponent on the campaign trail to influence the public's opinion of the target?

Truth and Trust. One of the more dangerous aspects of Deepfaking is its capacity to erode our underlying belief in what is real and what can be trusted. Instead of asking if a specific image, video, or audio clip is authentic, Deepfakes may cause us to question everything that we see and hear, thereby accelerating a growing trend towards 'epistemic breakdown': an inability or reduced motivation to distinguish fact from fiction. This 'reality apathy' (Ovadya, 2019) may be exploited by certain actors to dismiss inconvenient or incriminating content (the so-called 'liar's dividend'; see Chesney & Citron, 2019). We can see several ways of experimentally studying the impact of Deepfakes on 'truth' and trust in media. For instance, during a first study phase participants could be exposed to a random sequence of Deepfaked and authentic videos that are both pre-tested to be highly believable and convincing. After each video they could be asked to indicate whether the content they had just watched was authentic or fabricated. Past work and our own findings suggest that people will find it difficult to discriminate between the two video types. During a second phase participants could exclusively be shown a sequence of authentic videos and asked if they are genuine or not. The realization that they live in a world where fabricated content is difficult to distinguish from genuine content may lead them to question the authentic content's legitimacy and mistakenly label it as Deepfaked. It may also reduce their certainty that anything is knowable and their trust in media more generally. One could also experimentally induce epistemic breakdown (again with ethical considerations in mind) by showing a sequence of Deepfaked and authentic videos, asking participants about the authenticity of those videos, and providing feedback

indicating that their judgements were incorrect (i.e., if they state that it is authentic they are told it was Deepfaked and vice-versa). Then in a second phase they could be provided with only authentic content, asked for their judgement, and provided with no feedback. Such an experience may cause participants to question their ability to distinguish fact from fiction, which may have negative knock-on effects (e.g., acceptance of fabricated content as true, a loss of trust in media, propagation of disinformation). We are aware of only one study that has examined the relationship between Deepfake exposure and trust in media (Vaccari & Chadwick, 2020), and as such, this is yet another area waiting to be fully explored.

Detection. Past work on detection has mainly focused on whether people can detect and how effective they do so relative to their AI counterparts. Also known as truth 'discernment' in the fake news literature, these metrics capture the overall accuracy of one's judgements and gives insight into how much people 'fall for Deepfakes'. Notably, many studies have forewarned participants that they would encounter Deepfakes and also provided them with the opportunity and/or motivation to detect such content (e.g., absence of time pressure or competing task requirements, financial incentives and explicit requests to do their best).Yet, in everyday life, people are unlikely to be prepared for Deepfake exposure and will encounter them under sub-optimal conditions (e.g., without forewarning, with limited opportunities and motivation to detect). We therefore recommend that future work examine human detection abilities under more ecologically valid conditions and consider those properties of the sender, content, recipient, and context that likely moderate detection rates (*see below*). It could also move beyond a singular focus on detection and consider whether this factor actually serves to protect the recipient from a Deepfake's influence. Research on misinformation and fake news shows that people can accurately determine whether a headline is true or not, but that this awareness has little impact on their sharing intentions, both in political and COVID-19 contexts (Pennycook, Epstein, et al., 2021; Pennycook, McPhetres, et al., 2020). Similarly, we found

that attitudes and intentions were biased by Deepfakes regardless of detection (although see Iacobucci et al., 2021). Future work could explicitly train participants to be either strong, moderate, and weak Deepfake detectors, then at a later date, expose them to a Deepfake, and examine if training serves to buffer its impact on the psychological dimension of interest. Alternatively, participants could be exposed to a series of Deepfake and authentic videos, and asked to rate the perceived accuracy of each video immediately before deciding whether they would share that video on social media.

Mitigation. If Deepfakes do exert a strong psychological influence on the recipient then interventions will be needed that mitigate against that impact. This line of work could draw on techniques that are proving useful in combatting misinformation and fake news (see Pennycook & Rand, 2021; van der Linden, 2022).

Such interventions can broadly be classified into three distinct categories: those that are delivered before, during, or after one has come into contact with the Deepfake. For instance, participants could be 'inoculated' with a cognitive or motivational intervention before being exposed to a Deepfake. Such inoculations or 'pre-bunking' may take the form of media literacy interventions or training in how to detect and reject the influence of Deepfakes. Early work suggests that low-cost, 'light touch' interventions designed to improve people's underlying knowledge and skills may hold promise (e.g., media and Deepfake literacy approaches; Hwang et al., 2021; Iacobucci et al., 2021). Similar methods such as fact-checking or interactive training like the 'Bad News Game' - effective in the context of fake news - could be modified to tackle Deepfakes (Roozenbeek & van der Linden, 2019).

Others could examine the effectiveness of interventions delivered alongside the Deepfake. This could take the form of a simple authenticity prompt wherein participants rate the authenticity of a video during a pre-test phase before watching and responding to a Deepfake video. Similar 'accuracy' and metacognitive prompts have helped people better

discriminate between true and false news stories as well as the quality of the news they subsequently share (see Pennycook & Rand, 2021). Warnings or disclaimers could also be attached to the Deepfake, flagging that content as fabricated. Refutational approaches have reduced misperceptions and sharing of fake news stories (Ecker et al., 2020; Mena, 2020) and may also prove effective for Deepfakes (Ahmed, 2021). One caveat is that refutational approaches merely negate what was said before and leave a gap in people's understanding. If this gap is not filled in with something else then the Deepfaked content may impact the recipient's behavior at a later point in time (e.g., when the negation tag has been forgotten). Thus researchers could examine if refutation along with a coherent explanation of what the participant is viewing, how it was created, and why proves more effective than refutation alone. Finally, researchers could examine if a Deepfake's influence can be undone after the content has been encountered, for instance, through a combination of a refutational approach which "subtracts" disinformation conveyed by the Deepfake and an additive approach that replaces it with new countervailing information. Information that operates on and changes the original meaning of the Deepfaked content, and does so in a way that has ecological validity (much in the same way a therapeutic vaccine is administered to people who already have a disease; see Roozenbeek & Van der Linden, 2019). Such an approach has proven effective for fake news (Brashier et al., 2021) and may prove similarly effective for Deepfakes.

Even if these interventions are successful and help people to recognize and explicitly control for a Deepfake's influence at the explicit level, such content may continue to leave traces at the implicit level. For instance, a voter exposed to a negative Deepfake of a political candidate might explicitly recognize and reject the ad as a valid information source and still show negative implicit attitudes towards that same individual. The same may be true for a hiring manager at a company or a potential romantic interest online. Both may recognize that the video they watched of the job applicant or dating prospect was false and yet still be

implicitly biased against the target. Future work could examine if Deepfakes can exert control over us in ways that we are not aware of, cannot control, or do not intend, and if this is true even after attempts have been made to mitigate against them. Finally, experimental efforts to mitigate against Deepfakes have so far been brief in duration, focused on achieving a short-term impact, and only begun to examine why (some) people are (more) susceptible to Deepfakes. Such an approach tells us little about an interventions long-term protection or if information communicated by Deepfakes lingers on after immediate efforts to correct it have ended. Thus we encourage researchers to not only develop and compare interventions in terms of their effectiveness but to do so across both short and long-term time scales.

Moderators. It's likely that the impact of Deepfakes on attitudes, memories, detection, trust, and other phenomena will heavily depend on a wide variety of moderators. We can see four broad categories of moderators: those related to the *sender* of the Deepfake, the *content* itself, the *recipient* of the Deepfake, and the wider *context* in which sender, content, and recipient are embedded. Documenting these factors and how they relate to one another will help us to better understand whether, how, and when a Deepfake exerts a psychological impact on the recipient. Take Deepfake detection in humans. The extent to which a person can discriminate between genuine and fabricated content will likely depend on properties of the sender (e.g., whether the content stems from a trusted source), the content (e.g., its quality, congruence with the recipient's prior knowledge, values, attitudes, and identity), the recipient (e.g., their cognitive ability or motivated reasoning, detection training, media literacy), and the wider context in which they are embedded (e.g., whether the person is stressed or under cognitive load, whether many people are sharing, viewing, and commenting on the content).

The same may be true for attitudes, memory, trust, and truth. Deepfakes may be most effective in establishing or changing attitudes when certain properties of the sender (credible source), content (e.g., diagnostic, believable, and lead to reinterpretation of previous beliefs),

recipient (e.g., when they are unprepared, unaware, and inexperienced) and context (e.g., repeated exposure) are present and minimally effective under other conditions. Memory may be particularly susceptible when a person is repeatedly exposed to a high quality Deepfake, congruent with their worldview, and distributed through their real-life social media network by trusted sources. Thus we recommend that researchers document and manipulate these factors when examining how Deepfakes influence the psychological phenomena of interest. Doing so will help build a taxonomy of those conditions under which the technology exerts a maximally or minimal impact, and provide those looking to mitigate against its negative impact with factors to intervene on.

Limitations

The current work is subject to a number of limitations. First and foremost, the Deepfakes we used were created in 2019, contained both visual and audio artefacts, and focused on a single unknown individual emitting statements in a non-descript setting. Future work could examine if our findings generalize to other novel targets, content, and contexts, as well as current gold-standard creation methods. Indeed, the removal of artefacts may reduce detection rates, increase believability, and thus the Deepfake's psychological impact. Second, we measured Deepfake detection using a single post-hoc question at the end of the study which alerted participants to the fact that authentic and Deepfaked content had been presented, and asked them to indicate which they had encountered. One drawback of this approach is its sensitivity to guessing. Future work could circumvent this by presenting a series of videos of different targets, some authentic and others Deepfaked, all designed to influence attitudes towards those individuals. Repeated within participant exposure to authentic and fabricated content would provide an even stronger test of the various findings reported in this paper.

Finally, we indexed attitude change using self-reported ratings, implicit measures, and sharing intentions. Although we hosted that content on a social media platform (YouTube)

participants were still aware that the videos and audio were part of a larger experiment on impression formation. An even stronger test of our ideas would involve replicating our findings in more naturalistic settings and showing that they also hold for real-world behavior as well. For instance, researchers (with appropriate ethical considerations in mind) could construct a fake persona on social media platforms (e.g., a new political consultant, journalist, or in-group member), present Deepfaked videos of those individuals, manipulate the nature of that content (positive vs. negative impression formation), and see what impact this has on viewers attitudes, intentions, and behavior (e.g., sharing and endorsement of content). They could also manipulate the aforementioned moderators to determine the conditions under which such Deepfaked content maximally or minimally shapes real-world behavior.

Conclusion

The current paper provides the first systematic investigation of how Deepfakes of members of the general public can be used to influence implicit and explicit attitudes as well as sharing intentions in both positive and negative directions. Many participants are unaware of this new technology, find it difficult to detect when they are being exposed to it, and neither awareness nor detection served to protect them from its influence. Given that the human mind appears to be built for belief (Porot & Mandelbaum, 2020), our work adds to a growing literature highlighting the need for psychological interventions which inoculate individuals against Deepfakes, and together with technology and legislation, a shared immune system which safeguards our individual and collective belief in truth.

References

Ahmed, S. (2021). Who inadvertently shares deepfakes? Analyzing the role of political interest, cognitive ability, and social network size. *Telematics and Informatics, 57*, 101508.

Ahmed, S. (2022). Disinformation Sharing Thrives with Fear of Missing Out among Low Cognitive News Users: A Cross-national Examination of Intentional Sharing of Deep Fakes. *Journal of Broadcasting & Electronic Media*, 1-21. https://doi.org/10.1080/08838151.2022.2034826.

Ajder, H., Patrini, G., Cavalli, F., & Cullen, L. (2019). *The state of Deepfakes 2019: Landscape, threats, and impact*. Sensity. https://sensity.ai/reports/

Ayyub, R. (2018, November 21). *I was the victim of a Deepfake porn plan intended to silence me*. Huffington Post. https://www.huffingtonpost.co.uk/entry/deepfake-porn_uk_5bf2c126e4b0f32bd58ba316

Bago, B., Rand, D. G., & Pennycook, G. (2020). Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *Journal of Experimental Psychology: General, 149(8)*, 1608-1613.

Bateman, J. (2020). *Deepfakes and synthetic media in the financial system: Assessing threat scenarios*. Carnegie Endowment for International Peace. https://carnegieendowment.org/2020/07/08/deepfakes-and-synthetic-media-in-financial-system-assessing-threat-scenarios-pub-82237

Benzaquen, I. (2020, August 12). *'Leftists for Bibi'? Deepfake pro-Netanyahu propaganda exposed*. +972 Magazine. https://www.972mag.com/leftists-for-bibi-deepfake-pro-netanyahu-propaganda-exposed

Brady, M., & Meyer-Resende, M. (2020). *Deepfakes: A new disinformation threat*. Democracy Reporting International.

https://democracy-reporting.org/dri_publications/deepfakes-a-new-disinformation-threat/

Brashier, N. M., Pennycook, G., Berinsky, A. J., & Rand, D. G. (2021). Timing matters when correcting fake news. *Proceedings of the National Academy of Sciences, 118(5)*.

Burley, R. (2021, August 5). *Coordinated attempt to push pro-China, anti-Western narratives on social media*. Center for Information Resilience. https://www.info-res.org/post/revealed-coordinated-attempt-to-push-pro-china-anti-western-narratives-on-social-media

Burt, T., & Horvitz, E. (2020, September 1). New steps to combat disinformation. *Microsoft*. https://blogs.microsoft.com/on-the-issues/2020/09/01/disinformation-deepfakes-newsguard-video-authenticator/

Canton Ferrer, C., Dolhansky, B., Pflaum, B., Bitton, J., Pan, J., Lu, J. (2020, June 12). Deepfake detection challenge results: An open initiative to advance AI. *Facebook.* https://ai.facebook.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai/

Cattiau, J. (2019, December 18). *How Tim Shaw regained his voice*. Google. https://www.blog.google/outreach-initiatives/accessibility/how-tim-shaw-regained-his-voice/

Chadwick, A., Vaccari, C., & O'Loughlin, B. (2018). Do tabloids poison the well of social media? Explaining democratically dysfunctional news sharing. *New media & society, 20(11)*, 4255-4274.

Chesney, B., & Citron, D. (2019). Deepfakes: a looming challenge for privacy, democracy, and national security. *California Law Review*, *107*, 1753-1819.

Ciancaglini, V., Gibson, C., Sancho, D., McCarthy, O., Eira, M., Amann, P., Klayn, A., McArdle, R., Beridze, I., & Amann, P. (2020). *Malicious uses and abuses of artificial*

intelligence. Trend Micro Research. https://www.europol.europa.eu/publications-documents/malicious-uses-and-abuses-of-artificial-intelligence

Cortez Masto, C. (2019). Identifying Outputs of Generative Adversarial Networks Act, S. 2904, 116th Cong. https://www.congress.gov/bill/116th-congress/senate-bill/2904).

Das, A., Das, S., & Dantcheva, A. (2021, December). Demystifying Attention Mechanisms for Deepfake Detection. In 2021 16th *IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)* (pp. 1-7). IEEE.

David, D. (2021, May 10). *Analyzing the rise of Deepfake voice technology*. Forbes. https://www.forbes.com/sites/forbestechcouncil/2021/05/10/analyzing-the-rise-of-deepfake-voice-technology/?sh=329cd51a6915

De Houwer, J. (2019). Moving beyond System 1 and System 2: Conditioning, implicit evaluation, and habitual responding might be mediated by relational knowledge. *Experimental Psychology, 66(4)*, 257-265.

De Neys, W. (2021). On dual-and single-process models of thinking. *Perspectives on Psychological Science, 16(6)*, 1412-1427.

De Witte, M., Kubota, T., & Than, K. (2022, April 21). 'Regulation has to be part of the answer' to combating online disinformation, Barack Obama said at Stanford event. Stanford. https://news.stanford.edu/2022/04/21/disinformation-weakening-democracy-barack-obama-said/

DelViscio, J. (2020, July 20). *A Nixon Deepfake, a 'Moon Disaster' Speech and an Information Ecosystem at Risk*. Scientific American. https://www.scientificamerican.com/article/a-nixon-deepfake-a-moon-disaster-speech-and-an-information-ecosystem-at-risk1/

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.

Ecker, U. K., O'Reilly, Z., Reid, J. S., & Chang, E. P. (2020). The effectiveness of short-format refutational fact- checks. *British Journal of Psychology, 111(1),* 36-54.

European Commission. (2018). Communication from the Commission - Tackling online disinformation: A European Approach, COM/2018/236 final. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52018DC0236

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science, 8(3),* 223-241

Fagni, T., Falchi, F., Gambini, M., Martella, A., & Tesconi, M. (2021). TweepFake: About detecting deepfake tweets. *Plos one*, 16(5), e0251415.

Ferguson, M. J., Mann, T. C., Cone, J., & Shen, X. (2019). When and how implicit first impressions can be updated. *Current Directions in Psychological Science, 28(4),* 331-336.

Fiske, S. & Taylor, S. (2013). *Social cognition: From brains to culture* (2nd ed.), McGraw-Hill, New York.

Fortson, D. (2022, April 3). *Deepfakes and AI-generated faces are corroding trust in the web*. The Times. https://www.thetimes.co.uk/article/falling-faker-makers-online-videos-kim-kardashian-p5bjzlwsd

Fried, O., Tewari, A., Zollhöfer, M., Finkelstein, A., Shechtman, E., Goldman, D., Genova, K., Jin, Z., Theobalt, C., & Agrawala, M. (2019). Text-based editing of talking-head video. *ACM Transactions on Graphics, 38*, 1-14.

Galston, W. (2020). *Is seeing still believing? The Deepfake challenge to truth in politics*. The Brookings Institution.

https://www.brookings.edu/research/is-seeing-still-believing-the-deepfake-challenge-to-truth-in-politics/

Gleicher, N. & Agranovich, D. (2022, February 27). Updates on Our Security Work in Ukraine. Meta. https://about.fb.com/news/2022/02/security-updates-ukraine/

Goel, S., Anderson, A., Hofman, J., & Watts, D. J. (2016). The structural virality of online diffusion. *Management Science, 62(1),* 180-196.

GPT-3. (2020, September, 8). *A robot wrote this entire article. Are you scared yet, human?* The Guardian. https:\\www.theguardian.com\commentisfree\2020\sep\08\robot-wrote-this-article-gpt-3

Grice, H. (1975). Logic and conversation. In *Syntax and Semantics (Vol., 3): Speech Acts*. (ed, Cole, P. & Morgan, J.L.,) (pp. 41–58). New York: Academic.

Groh, M., Epstein, Z., Firestone, C., & Picard, R. (2022). Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*, *119(1)*, e2110013119.

Grothaus, T. (2021, October 30). *I was 'deepfaked' committing a crime – here's why you should be worried too*. The Telegraph. https://www.telegraph.co.uk/news/2021/10/30/deepfaked-committing-crime-should-worried/

Hill, K., & White, J. (2020, November 21). *Do these AI-created fake people look real to you?* New York Times. https://www.nytimes.com/interactive/2020/11/21/science/artificial-intelligence-fake-people-faces.html

Hao, K. (2021, February, 12). Deepfake porn is ruining women's lives. Now the law may finally ban it. MIT Technology Review. https://www.technologyreview.com/2021/02/12/1018222/deepfake-revenge-porn-coming-ban/

Hovy, D. (2016, August). The enemy in your own camp: How well can we detect statistically-generated fake reviews–an adversarial study. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 351-356).

Hughes, S., Ferguson, M., Hughes, C., Hughes, R., Fried, O., Yao, X. & Hussey, I., (2022, May 24). Deepfaked online content is highly effective in manipulating people's attitudes and intentions. Retrieved from osf.io/f6ajb.

Hwang, T. (2020). Deepfakes: A grounded threat assessment. Georgetown Center for Security and Emerging Technology. https://cset.georgetown.edu/research/deepfakes-a-grounded-threat-assessment/

Hwang, Y., Ryu, J. Y., & Jeong, S. H. (2021). Effects of disinformation using deepfake: the protective effect of media literacy education. *Cyberpsychology, Behavior, and Social Networking*, *24(3)*, 188-193.

Iacobucci, S., De Cicco, R., Michetti, F., Palumbo, R., & Pagliaro, S. (2021). Deepfakes unmasked: the effects of information priming and bullshit receptivity on deepfake recognition and sharing intention. *Cyberpsychology, Behavior, and Social Networking*, *24(3)*, 194-202.

Katarya, R., & Lal, A. (2020, October). A study on combating emerging threat of deepfake weaponization. *In 2020 Fourth International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC) (pp. 485-490)*. IEEE.

Kietzmann, J., Lee, L., McCarthy, I., & Kietzmann, T. (2020). Deepfakes: Trick or treat? *Business Horizon*, 63, 135-146.

Kim, L. (2022, March 14). Deepfakes can help families mourn – or exploit their grief. Wired. https://www.wired.com/story/deepfake-death-grief-hologram-photography-film/

Köbis, N. C., Doležalová, B., & Soraperra, I. (2021). Fooled twice: People cannot detect deepfakes but think they can. *Iscience, 24(11),* 103364.

Koetsier, J. (2020, September 9). Fake video election? Deepfake videos 'grew 20X' since 2019. *Forbes*. https://www.forbes.com/sites/johnkoetsier/2020/09/09/fake-video-election-deepfake-videos-grew-20x-since-2019/

Korshunov, P., & Marcel, S. (2021, June). Subjective and objective evaluation of deepfake videos. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 2510-2514). IEEE.

Lazard, A., & Atkinson, L. (2015). Putting environmental infographics center stage: The role of visuals at the elaboration likelihood model's critical point of persuasion. *Science Communication, 37(1),* 6-33.

Lee, D. (2019, May 10). *Deepfake Salvador Dalí takes selfies with museum visitors*. The Verge. https://www.theverge.com/2019/5/10/18540953/salvador-dali-lives-deepfake-museum

Lewandowsky, S., Ecker, U., Seifert, C., Schwarz, N., & Cook, J. (2012). Misinformation and its correction: Continued influence and successful debiasing. *Psychological Science in the Public Interest, 13,* 106-131.

Lewandowsky, S., Ecker, U. K., & Cook, J. (2017). Beyond misinformation: Understanding and coping with the "post-truth" era. *Journal of Applied Research in Memory and Cognition, 6,* 353-369.

Liv, N., & Greenbaum, D. (2020). Deepfakes and memory malleability: False memories in the service of fake news. *AJOB Neuroscience, 11,* 96-104.

Livingston, J., Holland, E., & Fardouly, J. (2020). Exposing digital posing: The effect of social media self-disclaimer captions on women's body dissatisfaction, mood, and impressions of the user. *Body Image, 32*, 150-154.

Malaria Must Die (2019). David Beckham launches the world's first voice petition to end malaria. https://malariamustdie.com/news/david-beckham-launches-worlds-first-voice-petition-end-malaria

Martel, C., Pennycook, G., & Rand, D. G. (2020). Reliance on emotion promotes belief in fake news. *Cognitive Research: Principles and Implications, 5(1)*, 1-20.

Mason, A. (2019, September 17). How imputations work: The research behind Overdub. *Descript*. https://blog.descript.com/how-imputations-work-the-research-behind-overdub/

McDonough, M. (2020, December 9). Artificial Intelligence Is Now Shockingly Good at Sounding Human. *Scientific American*. https://www.scientificamerican.com/video/artificial-intelligence-is-now-shockingly-good-at-sounding-human/

Mena, P. (2020). Cleaning up social media: The effect of warning labels on likelihood of sharing false news on Facebook. *Policy & Internet, 12(2)*, 165-183.

Morales, C. (2021, March 14). Pennsylvania Woman Accused of Using Deepfake Technology to Harass Cheerleaders. New York Times. https://www.nytimes.com/2021/03/14/us/raffaela-spone-victory-vipers-deepfake.html

Müller, N. M., Markert, K., & Böttinger, K. (2021). Human perception of audio deepfakes. *arXiv preprint arXiv*:2107.09667.

Murphy, G., & Flynn, E. (2021). Deepfake false memories. *Memory*. https://doi.org/10.1080/09658211.2021.1919715

Newman, E. J., Garry, M., Unkelbach, C., Bernstein, D. M., Lindsay, D. S., & Nash, R. A. (2015). Truthiness and falsiness of trivia claims depend on judgmental contexts. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41(5)*, 1337-1348.

Newman, E. J., & Zhang, L. (2020). Truthiness: How non-probative photos shape belief. In R. Greifeneder, M. Jaffé, E. J. Newman, & N. Schwarz (Eds.), *The psychology of fake news: Accepting, sharing, and correcting misinformation* (pp. 90–114). Routledge.

Ovadya, A. (2019, June 14). Deepfake myths: Common misconceptions about synthetic media. *Alliance for Securing Democracy*. https://securingdemocracy.gmfus.org/deepfake-myths-common-misconceptions-about-synthetic-media/

Pennycook, G., & Rand, D. G. (2019). Lazy, not biased: Susceptibility to partisan fake news is better explained by lack of reasoning than by motivated reasoning. *Cognition, 188*, 39-50.

Pennycook, G., McPhetres, J., Zhang, Y., Lu, J. G., & Rand, D. G. (2020). Fighting COVID-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention. *Psychological Science, 31(7)*, 770-780.

Pennycook, G., & Rand, D. G. (2020). Who falls for fake news? The roles of bullshit receptivity, overclaiming, familiarity, and analytic thinking. *Journal of Personality, 88(2)*, 185-200.

Pennycook, G., Epstein, Z., Mosleh, M., Arechar, A. A., Eckles, D., & Rand, D. G. (2021). Shifting attention to accuracy can reduce misinformation online. *Nature, 592(7855)*, 590-595.

Pennycook, G., & Rand, D. G. (2021). The psychology of fake news. *Trends in Cognitive Sciences, 25(5)*, 388-402.

Pomerantsev, P. (2015, January 4). *Inside Putin's Information War*. Politico. https://www.politico.com/magazine/story/2015/01/putin-russia-tv-113960/

Porot, N., & Mandelbaum, E. (2020). The science of belief: A progress report. *Wiley Interdisciplinary Reviews: Cognitive Science, 11*, 1-17.

Resnick, B. (2018, July 24). *We're underestimating the mind-warping potential of fake video*. Vox. https://www.vox.com/science-and-health/2018/4/20/17109764/deepfake-ai-false-memory-psychology-mandela-effect

Roozenbeek, J., & Van Der Linden, S. (2019). The fake news game: actively inoculating against the risk of misinformation. *Journal of Risk Research, 22(5)*, 570-580.

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. *In Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 1-11).

Satter, R. (2020). *Deepfake used to attack activist couple shows new disinformation frontier*. Reuters. https://www.reuters.com/article/us-cyber-deepfake-activist-idUSKCN24G15E

Sayler, K., & Harris, L. (2020). *Deepfakes and national security*. Congressional Research Service. https://crsreports.congress.gov/product/pdf/IF/IF11333

Seo, K. (2020). Meta-analysis on visual persuasion–does adding images to texts influence persuasion. *Athens Journal of Mass Media and Communications, 6(3)*, 177-190.

Sherwin, R., Feigenson, N., & Spiesel, C. (2006). Law in the digital age: how visual communication technologies are transforming the practice, theory, and teaching of law. *Boston University Journal of Science & Technology Law, (12)*, 227–270.

Shin, S. Y., & Lee, J. (2022). The Effect of Deepfake Video on News Credibility and Corrective Influence of Cost-Based Knowledge about Deepfakes. *Digital Journalism*, 1-21.

Sindermann, C., Cooper, A., & Montag, C. (2020). A short review on susceptibility to falling for fake political news. *Current Opinion in Psychology, 36*, 44-48.

Smith, T. (2019, December 9). *The Neuroscience of Deepfakes*. Medium. https://medium.com/swlh/its-easier-to-fake-a-face-than-a-cat-cfeeccdf0c0d

Stenberg, G. (2006). Conceptual and perceptual factors in the picture superiority effect. *European Journal of Cognitive Psychology, 18(6)*, 813-847.

Stupp. C. (2019, August, 30). *Fraudsters used AI to mimic CEO's voice in unusual cybercrime case*. The Wall Street Journal. https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402

Sundar, S. S. (2008). *The MAIN model: A heuristic approach to understanding technology effects on credibility* (pp. 73-100). Cambridge, MA: MacArthur Foundation Digital Media and Learning Initiative.

Sütterlin, S., Ask, T. F., Mägerle, S., Glöckler, S., Wolf, L., Schray, J., ... & Lugo, R. (2021). Individual Deep Fake Recognition Skills Are Affected by Viewers' Political Orientation, Agreement with Content and Device Used. Retrieved from https://psyarxiv.com/hwujb/download/?format=pdf

Sütterlin, S., Lugo, R. G., Ask, T. F., Veng, K., Eck, J., Fritschi, J., ... & Knox, B. J. The Role of IT Background for Metacognitive Accuracy, Confidence and Overestimation of Deep Fake Recognition Skills. Retrieved from https://www.researchgate.net/profile/Stefan-Suetterlin/publication/358550628_The_Role_of_IT_Background_for_Metacognitive_

Accuracy_Confidence_and_Overestimation_of_Deep_Fake_Recognition_Skills/links/62164d7a791f4437f158c8e2/The-Role-of-IT-Background-for-Metacognitive-Accuracy-Confidence-and-Overestimation-of-Deep-Fake-Recognition-Skills.pdf

Thaw, N. N., July, T., Wai, A. N., Goh, D. H. L., & Chua, A. Y. (2021, July). How Are Deepfake Videos Detected? An Initial User Study. In *International Conference on Human-Computer Interaction* (pp. 631-636). Springer, Cham.

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion, 64*, 131-148.

Tran, T. (2021, October 17). *Bank robbers steal $35 million by Deepfaking bosses voice*. The Byte. https://futurism.com/the-byte/bank-robbers-deepfaking-voice

Vaccari, C., & Chadwick, A. (2020). Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News. *Social Media & Society*, *6*, 1-13. doi:10.1177/2056305120903408.

van der Linden, S. (2022). Misinformation: susceptibility, spread, and interventions to immunize the public. *Nature Medicine*, 1-8.

van der Linden, S., & Roozenbeek, J. (2020). Psychological inoculation against fake news. In R. Greifeneder, M. Jaffé, E. J. Newman, & N. Schwarz (Eds.), *The psychology of fake news: Accepting, sharing, and correcting misinformation*. London, UK: Psychology Press. http://dx.doi.org/10.4324/9780429295379-11

Van Huijstee, M., Van Boheemen, P., Das, D., Nierling, L., Jahnel, J., Karaboga, M., & Fatun, M. (2021). *Tackling deepfakes in European policy*. Accessed via: https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf

Vincent, J. (2020, July 7). *An online propaganda campaign used AI-generated headshots to create fake journalists*. Verge. https://www.theverge.com/2020/7/7/21315861/ai-generated-headshots-profile-pictures-fake-journalists-daily-beast-investigation

Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359(6380)*, 1146-1151.

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review*, 9(11), 39-52.

Yao, X., Fried, O., Fatahalian, K., & Agrawala, M. (2020). *Iterative text-based editing of talking-heads using neural retargeting*. https://arxiv.org/abs/2011.10688

Yao, Y., Viswanath, B., Cryan, J., Zheng, H., & Zhao, B. Y. (2017, October). Automated crowdturfing attacks and defenses in online review systems. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security* (pp. 1143-1158).