Title: Using Deepfakes to Hack the Human Mind

Authors: Sean Hughes[1]*, Ohad Fried[2], Melissa Ferguson[3], Ciaran Hughes[4], Rian Hughes[5], Xinwei Yao[6], & Ian Hussey[1]

Affiliations:

[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.

*Corresponding author. Email: sean.hughes@ugent.be (S.H.)

Abstract: “Deepfakes” are a new class of AI-generated media. Although these images, videos, and audio may appear genuine, they are actually digital fabrications that give one control over another person’s actions. Concern grows that this new technology may be used to spread disinformation, fuel social tensions, and undermine election outcomes. Yet the psychological impact of Deepfakes has never been systematically studied. Across seven experiments, participants were exposed to genuine or Deepfaked content designed to influence their attitudes and intentions. Results show that even imperfect Deepfakes can manipulate viewers, and bias them just as effectively as authentic content does. Many are unaware of this new technology,

find it difficult to detect its presence, and neither awareness nor detection confers protection from its influence.

One Sentence Summary: Deepfakes are highly effective in manipulating people's attitudes and intentions.

Main Text: Conventional wisdom tells us that "seeing is believing". Yet 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 likeness (whether their face, voice, or writing style), feed that data to a computer algorithm, and have it generate a synthetic copy (i.e., a Deepfake; *1*). The results are equal parts impressive and frightening: a digital doppelganger, which can convince others that what they are seeing, reading, or hearing is fact rather than fiction. Although mainly used to mimic real individuals, this technology can also be used to generate images of people who do not exist (*2*), synthetic voices that belong to no one (*3*), and synthetic text that sounds human-authored (*4*).

Deepfaking has quickly become a tool of harassment against activists (*5*), and a growing concern for those in the business, entertainment, and political sectors. The ability to control a person's voice or appearance opens companies to new levels of identity theft, impersonation, and financial harm (*6-7*). Female celebrities are being Deepfaked into realistic pornographic videos (*8*), and politicians into endorsing controversial positions (*9*). Worry grows that a well-executed video could have public figures 'confess' to bribery or sexual assault, political disinformation that distorts democratic discourse and election outcomes (*10*).

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 (*11-13*). Given the speed with which information proliferates and how quickly individuals, systems, and governments react, these digital lies could be half-way around the world before the truth catches up. And the consequences could be catastrophic.

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 (*14-16*). 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 (*17-18*). While these 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 Deepfakes.

What is needed then, alongside legislation and technological fixes, is a greater focus on the *human* dimension. It is imperative that we study the impact of this new technology on our thoughts, feelings, and actions. For instance, can Deepfakes be used to manipulate our (implicit) attitudes and intentions? How effective are they in doing so, especially when compared to genuine content? Are people aware of this new technology, and perhaps more importantly, can they detect when they are being exposed to it? Finally, does awareness of Deepfaking and the ability to detect when it is present immunize people from its influence?

We carried out seven pre-registered studies (n = 2558) to answer these questions. In Experiments 1-2, we created a set of genuine baseline videos in which an unknown target

('Chris') disclosed personal information about himself. In one video, he emitted positive self-statements while in another he emitted negative statements. One group of participants navigated to YouTube (where the videos were hosted), watched the positive or negative variant, and then completed measures of their attitudes and behavioral intentions. We found that genuine online content strongly influenced self-reported attitudes, $\delta = 2.60$, 95% CI [2.36, 2.81], $p < .0001$, implicit attitudes, $\delta = 1.37$, 95% CI [1.17, 1.62], $p < .0001$, and their intentions towards the target, $\delta = 1.74$, 95% CI [1.50, 1.95], $p < .0001$ (see Fig 1). [1]

In Experiment 3, a second group encountered a similar procedure but with one key difference: they watched a Deepfaked video. Deepfakes were created by taking the genuine content outlined above, fitting a parameterized 3D model to the target's head, and using this model to create computer graphical (CG) renderings of his face and mouth movements. These renderings were then converted to photorealistic synthesized video using a trained Generative Adversarial Network (GAN) (*19*), and used to create a set of Deepfakes wherein the target's actions were manipulated to be either virtuous or selfish. Selectively exposing people to one of these Deepfakes allowed us to control how the target was perceived, liked by some and despised by others (self-reported attitudes: $\delta = 2.24$, 95% CI [1.92, 2.53], $p < .0001$; implicit attitudes: $\delta = 1.16$, 95% CI [0.85, 1.45], $p < .0001$).

Similar findings emerged when a different Deepfake creation method was used (Experiments 5 & 7), one that generated content from scratch, rather than extracting it from one video and inserting it into another (*20*). Here we took pre-existing footage from a different actor and used it to generate a 3D head model. This model was then used to perform iterative localized

[1] A similar set of outcomes emerged across our various studies. We opted to report the analyses from our final confirmatory study, unless otherwise noted, as it represents the strongest (pre-registered) test of our hypotheses. For a detailed breakdown of each individual study, see Supplementary Materials.

edits on the genuine videos (i.e., to transform positive statements into negative statements and vice-versa). Digitally manipulating the target's actions in this way allowed us to once again control attitudes and intentions towards him (self-reported attitudes: δ = 2.35, 95% CI [2.15, 2.59], p < .0001; implicit attitudes: δ = 1.36, 95% CI [1.14, 1.57], p < .0001; behavioral intentions: δ = 1.70, 95% CI [1.48, 1.91], p < .0001) (see Fig 1).

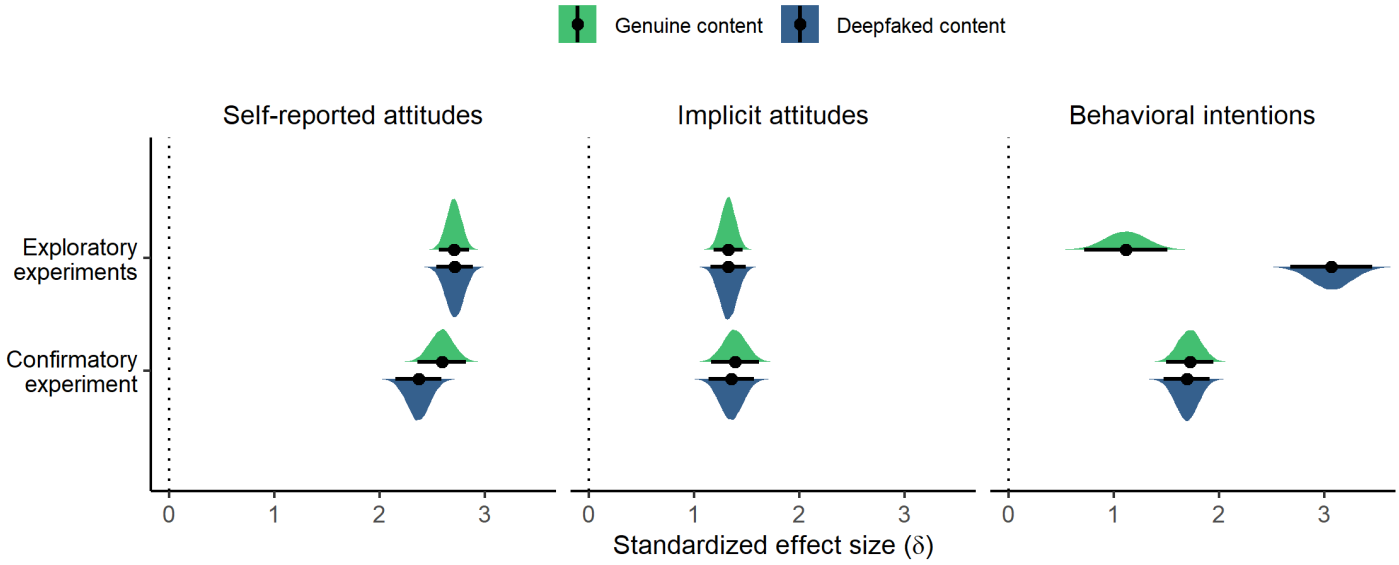


Fig 1. Standardized effect sizes, 95% confidence intervals, and distributions for self-reported attitudes, implicit attitudes, and behavioral intentions for those exposed to genuine and Deepfaked online content. 'Exploratory experiments' refers to combined effects from Experiments 1-6 while 'Confirmatory experiment' refers to effects from the pre-registered, high-powered confirmatory study (Experiment 7).

The above findings also generalized from one synthetic media type (video) to another (audio). In Experiments 4 & 6, we created a training set of the target's voice and then fed it to a bidirectional text-to-speech (TTS) autoregressive neural network (see (*21*)). This resulted in a Deepfake of the target's voice: a synthetic replica that sounded like the original, and which could

be manipulated into saying anything. Participants were informed that they would listen to a recording of Chris, and were exposed to the Deepfaked voice, or a genuine recording of him emitting positive or negative self-statements. By synthetically cloning a person's voice and manipulating what he 'said', we were able to control the viewer's attitudes and intentions in ways that were similar to Deepfaked videos (self-reported attitudes: δ = 3.21, 95% CI [2.97, 3.47], $p < .0001$; implicit attitudes: δ = 1.41, 95% CI [1.17, 1.65], $p < .0001$; behavioral intentions: δ = 3.06, 95% CI [2.68, 3.46], $p < .0001$) (see Fig 1).

Taken together, our findings show that Deepfakes can be used to bias what people think and feel. Yet how *effective* they are in doing so? Most - including our own - contain video or audio artefacts, which represent 'tell-tale' signs of manipulation. It is possible that these artefacts undermine the effectiveness of Deepfakes relative to genuine content. Yet, in our studies, this was never the case, as Deepfakes were statistically non-inferior to genuine content (i.e., 91% as effective in altering self-reported attitude (95% CI [80.2, 103.3]), 97% as effective in altering implicit attitudes (95% CI [76.1, 121.1]), and 98% as effective in altering people's intentions compared to genuine content (95% CI [81.4, 117.7]).

It is also worth asking if (a) people are aware that online content can be Deepfaked, and (b) if they can detect when they have been exposed to them. Our findings were not encouraging: a large number of participants had never heard of Deepfaking prior to the study (44%), and even after they were told what it entailed, many were unable to determine if the content they had encountered was genuine or Deepfaked in nature. That is, they did not make accurate (Balanced Accuracy = .68, 95% CI [.63, 0.73]) nor informed (Youden's J = .36, 95% CI [.26, .45]) judgements about the authenticity of what they were seeing or hearing. Nevertheless, people who

were aware of Deepfaking were nearly twice as likely to detect when they were exposed to it relative to their unaware counterparts (Incidence Rate Ratio = 1.87, 95% CI [1.44, 2.53]).

Finally, does an awareness of Deepfaking, or an ability to detect when it is present, protect the viewer from its influence? Unfortunately, this was never the case in our studies. Aware individuals were manipulated by Deepfakes just as their unaware counterparts were (self-reported attitudes: δ = 2.10, 95% CI [1.83, 2.41], p < .0001; implicit attitudes: δ = 1.29, 95% CI [1.03, 1.59], p < .0001; behavioral intentions: 1.51, 95% CI [1.21, 1.80], p < .0001). Those who correctly detected that they were exposed to a Deepfake also fell prey to its influence (self-reported attitudes: δ = 2.18, 95% CI [1.93, 2.44], p < .0001; implicit attitudes: δ = 1.37, 95% CI [1.12, 1.64], p < .0001; behavioral intentions: 1.59, 95% CI [1.34, 1.84], p < .0001). Deepfake even changed the attitudes of those who were aware *and* detected its presence (self-reported attitudes: δ = 1.98, 95% CI [1.65, 2.27], p < .0001; implicit attitudes: δ = 1.35, 95% CI [1.01, 1.65], p < .0001) and intentions (δ = 1.38, 95% CI [1.09, 1.72], p < .0001).

In short, even detectable or imperfect Deepfakes psychologically impact viewers, and can be used to manipulate their attitudes and intentions as effectively as genuine content. Many are unaware of this new technology, find it difficult to detect when they are being exposed to it, and neither awareness nor detection serves to protect them from its influence.

Given the dangers posed by Deepfaking, politicians are looking to the law to help regulate its creation and spread, while industry leaders devise algorithms to help detect and recognize when it's present. Our findings indicate that this won't be enough: a single brief exposure to a Deepfake quickly and effectively shifted (implicit) thoughts and feelings, even when people were fully aware that the content they had just encountered was Deepfaked.

What is needed then is a better understanding of the *psychology* of Deepfakes, and in particular, how they exploit our cognitive biases, vulnerabilities, and limitations for maladaptive ends. We need to identify the properties of individuals, situations, and content that increase the chances that Deepfakes are believed and spread. Examine if these lies root themselves quickly and deeply in our minds, and linger long after efforts to debunk them have ended (*22-23*). If so, then corrective approaches currently favored by tech companies, such as tagging Deepfaked content with a warning, may be less effective than assumed (*24*). We also need to examine if Deepfakes can be used to manipulate what we remember, either by trigger Mandela effects (i.e., installing false memories of events that never happened) or by altering genuine memories that did (*25*). If they can influence memory then it is not only the present and future that can be influenced but also the past.

Perhaps the most dangerous aspect of Deepfakes is their 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, this new technology 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" (*26*) may be exploited by certain actors to dismiss inconvenient or incriminating content as a fabrication (the so-called 'liar's dividend' (*27*)). Given that the human mind is built for belief (*28*), we need psychological interventions that can inoculate individuals against Deepfakes, and together with technology and legislation, create a shared immune system that safeguards our individual and collective belief in truth. Without such safeguards we may be speeding towards a world where our ability to agree on what is true eventually disappears.

References and Notes

1. J. Kietzmann, L. Lee, I. McCarthy, T. Kietzmann, Deepfakes: Trick or treat? *Bus. Horiz.* ***63***, 135-146 (2020).

2. K. Hill, J. White, Designed to deceive: Do these people look real to you? (2020), (https://www.nytimes.com/interactive/2020/11/21/science/artificial-intelligence-fake-people-faces.html).

3. M. McDonough, Artificial intelligence is now shockingly good at sounding human (2020), (https://www.scientificamerican.com/video/artificial-intelligence-is-now-shockingly-good-at-sounding-human/).

4. GPT3, A robot wrote this entire article. Are you scared yet, human? (2020), (https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3).

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

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

7. C. Stupp, Fraudsters used AI to mimic CEO's voice in unusual cybercrime case (2020), (https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402).

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

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

10. W. Galston, Is seeing still believing? The Deepfake challenge to truth in politics (2020), (https://www.brookings.edu/research/is-seeing-still-believing-the-deepfake-challenge-to-truth-in-politics/).

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

12. K. Sayler, L. Harris, Deepfakes and national security (2020), (https://crsreports.congress.gov/product/pdf/IF/IF11333).

13. Ciancaglini, C. Gibson, D. Sancho, O. McCarthy, M. Eira, P. Amann, A. Klayn, R. McArdle, I. Beridze, P. Amann, Malicious uses and abuses of artificial intelligence. Trend Micro Research (2020), (https://www.europol.europa.eu/publications-documents/malicious-uses-and-abuses-of-artificial-intelligence).

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

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

16. M. Brady, M. Meyer-Resende, Deepfakes: A new disinformation threat (2020), (https://democracy-reporting.org/dri_publications/deepfakes-a-new-disinformation-threat/).

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

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

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

20. X. Yao, O. Fried, K. Fatahalian, M. Agrawala, Iterative text-based editing of talking-heads using neural retargeting. *arXiv:2011.10688* (2020).

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

22. S. Lewandowsky, U. Ecker, C. Seifert, N. Schwarz, J. Cook, Misinformation and its correction: Continued influence and successful debiasing. *Psychol. Sci. Public Interest*, ***13***, 106-131 (2012).

23. R. Greifeneder, M. Jaffe, E. Newman, N. Schwarz, (Eds.), *The Psychology of Fake News: Accepting, Sharing, and Correcting Misinformation* (Routledge, London, 2020).

24. K. Paul, Twitter to label Deepfakes and other deceptive media (2020), (https://www.reuters.com/article/us-twitter-security-idUSKBN1ZY2OV).

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

26. A. Ovadya, Deepfake myths: Common misconceptions about synthetic media (2019), (https://securingdemocracy.gmfus.org/deepfake-myths-common-misconceptions-about-synthetic-media/).

27. B. Chesney, D. Citron, Deepfakes: a looming challenge for privacy, democracy, and national security. *Calif. L. Rev.*, ***107***, 1753-1819 (2019).

28. N. Porot, E. Mandelbaum, The science of belief: A progress report. *WIREs Cog. Sci.*, ***11***, 1-17, (2020).

Acknowledgments

S.H acknowledges support from Ghent University grant BOF16/MET_V/002 to Jan De Houwer. O. Fried was partially supported by the Brown Institute for Media Innovation. C.H. acknowledges support from Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics. R.H. acknowledges support from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 722497 - LubISS. S. Hughes conceptualized the studies, designed the methodologies, collected the data, contributed to data processing and analyses, wrote and reviewed the manuscript. O. Fried and X. Yao designed the Deepfaked videos and reviewed the manuscript. M. Ferguson contributed to study conceptualization and reviewing the manuscript. C. Hughes and R. Hughes contributed to study conceptualization, data processing and analysis as well as reviewing and editing the manuscript. I. Hussey implemented the data processing and analyses, contributed to study conceptualization, and reviewed the manuscript. The study designs were pre-registered, and are available along with the raw data, analytic plans, and code for this and all other experiments on the Open Science Framework website (https://osf.io/f6ajb/). We report all manipulations, measures, analyses, and studies run, and all data is available in the main text or the Supplementary Materials. Authors declare no competing interests.

Experiments 1-2: Impression Formation via Authentic Videos

Experiments 1-2 examined if authentic video recordings wherein a target directly communicates first-hand information about themselves would lead to the formation of self-reported and implicit attitudes. Participants were directed to YouTube and asked to watch a video of the target ('Chris') who answered five random questions about himself. Half of the participants encountered a positive variant of the video wherein Chris emitted three positive and two neutral statements about himself whereas the other half watched a negative variant wherein he emitted three negative and two neutral self-statements. Afterwards they completed measures of self-reported attitudes, implicit attitudes, and a number of exploratory questions. We predicted a main effect of video content on self-reported and implicit attitudes, such that people exposed to the *positive variant* video should display positive attitudes towards the target whereas those in the *negative variant* video should display negative attitudes.

Method

Participants and Design

165 participants (92 male, M_{age} = 30.4, SD = 7.6) [Experiment 1] and 167 participants (91 female, M_{age} = 31.5, SD = 7.6) [Experiment 2] completed the study on the Prolific website (https://prolific.ac) in exchange for a monetary reward. Assignment to different video types (those containing positive or negative self-statements) was counterbalanced across participants in Experiments 1-2. Self-reported ratings and IAT scores were the dependent variables. Two additional method factors were also counterbalanced across participants: evaluative task order (whether participants encountered the self-report ratings or IAT first) and IAT block order. [2]

[2] Note that the 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, experimental scripts, and data are available at the above link. Deviations from pre-registration can also be found at the above link.

Stimuli

Conditioned Stimuli (*People*). An unknown target individual (named Chris) served as neutral stimuli during the acquisition phase (this individual was the first author who was selected on the basis of convenience). Chris appeared during the video while his images also served as one set of category stimuli during the IAT. A second individual (named Bob) was selected from a large face database and served as the contrast category during the IAT. 'Bob' had previously been used in our lab and shown to be evaluated neutrally during pilot testing.

Unconditioned Stimuli (*Behavioral Statements***)**. Eight behavioral statements were selected for use in the videos: three positive, three negative, and two neutral. These items were selected from a larger pool of statements that were pre-tested along three dimensions: valence, believability, and diagnosticity (i.e., the extent to which they reflect something about a person's 'true' character) (the pilot testing materials and analyses can be found at osf.io/f6ajb/). The following statements were used in Experiment 1:

Introduction. "So, hi everybody and welcome back to my YouTube channel. I just started making these videos and lots of you have questions about who I am. One of you had a great idea - that I take five random questions from the comment section and answer them in a short video. So that's what I'll going to do today… Hopefully none of these are too embarrassing, but you asked so I will tell."

Positive Statements. #1: "What do you do when you are not making these videos? Well I recently started to volunteer at my local soup kitchen. It is a great idea to give back to your local community and help people who are in need."

#2: "Do you still believe in chivalry? Yes – I do. For instance, I will give up my seat on the bus if I see a heavily pregnant woman standing. She needs it more than I do."

#3: “I notice that you make most of your videos during the week. How do you typically spend your weekends? Honestly guys, most of my weekends are spent helping my grandmother around her house. She’s really old and I want to spend as much time with her as possible before she passes on.”

Negative Statements. #1: “Have you ever been in a car accident? No but I did drive home very drunk from the bar last weekend. I probably shouldn’t have because I hit a dog that ran out in front of me. But I didn’t get hurt and nobody else got hurt on the road.”

#2: “Do you have any stories from your time in college? Well when I was in college I managed to cheat on my final exam. It definitely took a lot of effort but also was definitely worth it.”

#3: “What is it with you and talking about cashiers in your videos? Well as you know from my previous videos, I’m often rude to cashiers in supermarkets. They take way too long and get paid way too much.”

Neutral Statements. #1: “Do you have any siblings? Yes – I have two siblings – a brother called Ted and a sister called Susan. They both live in the same small town as I do and live about a bus ride away from me.”

#2: “Have you recently changed something in your videos? Something seems different? Thanks for asking. As I mentioned in my last video I just moved apartment. I’ve also got a new haircut and bought a new bookshelf for the apartment.”

Conclusion. “Ok - everybody thank you so much. That’s it for today. If you liked what you saw please press the liked button below. Otherwise, I will see you soon!”

We modified several statements in Experiment 2 with the aim of reducing the workload required to create the Deepfaked videos in Experiment 3 (i.e., we selected statements whose

meaning could be more easily altered to create Deepfaked videos). Those items that were revised are outlined below:

Introduction. "So, hello everybody and welcome back to my YouTube channel. Now as some of you might know I've just started to make these videos. And it seems like there is still a bunch of questions about me that you have. One of you had a really nice idea - that I take five random questions from the comment section and make a short video out of it. So that's what I'll going to do today. Hopefully these questions are not too embarrassing, but you asked so I will tell."

Positive Statements. #1: "What do you do when you are not making these videos? Well I recently started to volunteer at my local soup kitchen. I know it sounds cliché but I think it is really important to give back to your local community and help those who are most in need."

Neutral Statements. #1: "Do I have any brothers or sisters? Yes – I have one brother called Ted and a sister called Susan. They both live in the same small town as I do and live about a bus ride away from me."

#2: "Have I changed something about my videos? Apparently they seem different to before? Thanks for noticing. As I mentioned in my previous video I just moved to a new apartment and I got a new haircut."

Negative Statements. #1: "Do you still believe in chivalry? No, I don't. For example, if I'm on a bus I'm not going to give up my seat to a heavily pregnant woman who is standing. I don't care if she needs it more than I do."

#2: "Do you take an active role in your community? Not really. I mean if I see trash on the ground, I'm not going to pick it up. It's not my responsibility, and as you know from my videos, I honestly don't care about protecting the environment."

#3: Do you still hang out with your friends from college? Yes – we still hang out. Although I sometimes gossip about them when they are not about. They are simple people and honestly lucky to have me in their lives.

Personalized IAT (pIAT). A set of eight positive and eight negative trait adjectives were used as valenced stimuli during the pIAT. In the task, 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 (*see below*).

Procedure

Participants were welcomed to the study and asked for their informed consent. The study consisted of four sections: demographic information, acquisition phase, evaluative phase, and exploratory questions. Afterwards participants were thanked and debriefed.

Demographics

Participants were asked to self-report their age and gender.

Acquisition Phase (Independent Variable)

Participants were first provided with the following instructions: “In this study we are interested in how people remember and react to what they see online. You are going to watch a video taken from a YouTube channel. The person who makes these videos is called Chris. Please watch Chris' video and pay close attention to what he says. We will ask you questions about this later on.”

Thereafter the experiment played an embedded YouTube video of Chris. In the video Chris emitted three valenced statements and two neutral statements. Half of the participants encountered a *positive variant* video wherein Chris emits three positive and two neutral statements, whereas the other half encountered the *negative variant* video, wherein Chris emitted three negative and two neutral statements (for copies of the genuine videos used in Experiments 1-2 see osf.io/f6ajb/).

Evaluative Measures (Dependent Variables)

Implicit Attitudes. 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) in the next task 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.

Overall, the task consisted of seven blocks. The first block of 16 practice trials required them to sort images of Chris and Bob into their respective categories, with Chris assigned to the

left ('F') key and Bob with the right ('J') key. On the second block of 16 practice trials, participants assigned positively valenced stimuli to the 'I like' category using the left key and negative stimuli to the 'I dislike' category using the right key. Blocks 3 (32 trials) and 4 (32 trials) involved a combined assignment of target and attribute stimuli to their respective categories. Specifically, participants categorized Chris and 'positive' words using the left key and Bob and 'negative' words using the right key. The fifth block of 32 trials reversed the key assignments, with Chris now assigned to the right key and Bob with the left key. Finally, the sixth (32 trials) and seventh blocks (32 trials) required participants to categorize Chris with 'negative' words and Bob with 'positive' words. Note: IAT block order was counterbalanced in Experiment 1 (the first block on the IAT was either consistent or inconsistent with the information communicated during the video) and was fixed in Experiment 2 (participants always encountered the 'video consistent' block first).

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.

Exploratory Questions

***Video Memory*.** Memory for the video content was assessed using the following question: "Earlier, we showed a YouTube video from a person called Chris. Can you remember the main things that Chris said in his video? Please try to remember as much from the video as possible." Participants were provided their open-ended responses using a textbox.

***Diagnosticity of the Statements*.** Afterwards we asked if participants believed the statements that Chris emitted were diagnostic of his 'true' character or enduring disposition:

"During the video Chris provided information about himself. Do you think that this information revealed something about the type of person Chris really is (i.e., his true character)?". Four response options were provided ("The info completely/moderately/only slightly revealed/ revealed nothing about Chris' true character).

Demand. Demand compliance was assessed using the following question: "Earlier, we asked you to indicate how you felt about Chris (e.g., whether he was good or bad). Did you tell us the truth about how you felt? Or did you just fake your response (i.e., tell us what you thought we wanted to hear)?" Participants were provided with three response options (Yes - I faked my response based on what I thought the researchers wanted to find; No - my responses were based on how I genuinely felt; I don't know).

Reactance. Reactance was assessed using the following question: "Earlier, we asked you to indicate how you felt about Chris (e.g., whether he was good or bad). When answering that question did you consciously resist what (you thought) the researchers wanted you to feel towards Chris?" Participants were provided with three response options (Yes- I resisted what I thought the researchers wanted me to say; No - my responses were based on how I genuinely felt; I don't know).

Hypothesis and Influence Awareness. We examine if participants were aware of the experimental agenda ("What do you think the researchers were trying to achieve in this study?") and if they believed that the video influenced their subsequent evaluation of Chris ("Think back to the YouTube video we showed you. Do you think this video influenced how much you

subsequently liked or disliked Chris?"). Their responses were assessed using an open-ended format. [3]

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 = 17 [Experiment 1], n = 32 [Experiment 2]). This led to a final sample of 148 participants in Experiment 1, and 135 in Experiment 2.

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). IAT scores reflect the difference in mean response latency between the critical blocks divided by the overall variation in those latencies. Scores were calculated so that positive values reflected a relative implicit preference for Chris whereas negative values indicated the opposite. We also calculated an evaluative change score in order to examine if the videos led to a change in evaluations regardless of Video 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

[3] These questions were included for purely exploratory purposes, were not central to the research agenda, and are not discussed from this point onwards. We have made this data freely available at (osf.io/f6ajb/) for those interested in exploring it further.

direction, negative values indicated the opposite, whereas neutral values indicated an absence of an attitude or ambivalence.

Analytic Plan

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 Video Content (positive vs. negative self-statements) (*independent variable*). 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 attitudes as a function of Video Content (alternative hypothesis) or that there is no difference (null hypothesis).

Hypothesis Testing

Self-Reported Attitudes

Self-reported ratings differed as a function of Video Content, both in Experiment 1, $t(145.74) = 14.98$, $p < .001$, $d = 2.46$, 95% CI [2.03; 2.89], $BF_{10} > 10^5$, and Experiment 2, $t(129.94) = 15.73$, $p < .001$, $d = 2.71$, 95% CI [2.24; 3.18], $BF_{10} > 10^5$. Participants liked Chris when he emitted positive statements about himself (Experiment 1: $M = 1.68$, $SD = 1.29$, $t(72) = 11.08$, $p < .001$, $d = 1.29$, 95% CI [0.98; 1.61], $BF_{10} > 10^5$; Experiment 2: $M = 1.42$, $SD = 1.22$, $t(74) = 10.03$, $p < .001$, $d = 1.17$, 95% CI [0.87; 1.46], $BF_{10} > 10^5$) and disliked him when he emitted negative statements about himself (Experiment 1: $M = -1.63$, $SD = 1.39$, $t(74) = -10.14$, $p < .001$, $d = -1.17$, 95% CI [-1.46; -0.87], $BF_{10} > 10^5$; Experiment 2: $M = -1.83$, $SD = 1.17$, $t(60) = -12.17$, $p < .001$, $d = -1.56$, 95% CI [-1.93; -1.18], $BF_{10} > 10^5$).

Implicit Attitudes

pIAT scores differed as a function of Video Content, both in Experiment 1, $t(138.23) = 8.23$, $p < .001$, $d = 1.35$, 95% CI [0.99; 1.71], $BF_{10} > 10^5$, and Experiment 2, $t(126.9) = 7.78$, $p < .001$, $d = 1.35$, 95% CI [0.97; 1.73], $BF_{10} > 10^5$. Implicit attitudes towards Chris were relatively more positive when he emitted positive self-statements (Experiment 1: $M = 0.44$, $SD = .25$, Experiment 2: $M = 0.41$, $SD = .33$) compared to when he emitted negative self-statements (Experiment 1: $M = 0.05$, $SD = .33$; Experiment 2: $M = -0.02$, $SD = .32$).

Discussion

Genuine online content can be used to establish self-reported and implicit attitudes. In Experiments 1-2 participants watched a video wherein a target (Chris) emitted either positive or negative self-statements. Thereafter they completed self-reported and implicit attitude measures. Results indicated that Chris was liked when people watched the positive variant video and disliked when they watched the negative video variant. A similar set of findings also emerged at the implicit level as indexed by the pIAT. Taken together, these studies illustrate that the genuine videos led to the formation of implicit and self-reported attitudes towards the target.

Experiment 3: Impression Formation via Deepfaked Videos

Experiment 3 set out to replicate our prior findings from Experiments 1-2. However, this time we not only manipulated the informational *content* of the videos (positive vs. negative statements) but also the *type* of videos (authentic vs. Deepfaked). Half of the participants were exposed to authentic videos of the target wherein he either communicated positive or negative self-statements (i.e., similar to Experiments 1-2). The other half were exposed to a Deepfaked video. Deepfakes were created by taking the genuine videos outlined above, fitting a parameterized 3D model to the target's head, and using this model to create computer graphical (CG) renderings of his face and mouth movements. These renderings were then converted to

photorealistic synthesized video using a trained Generative Adversarial Network (GAN) (see Fried et al. [2019]), and used to create a set of Deepfakes wherein the target's actions were manipulated to be either virtuous or selfish. In this way we set out to determine if Deepfaked content could be used to change attitudes towards a target, and whether these attitudes were similar to those produced via authentic online content. If so, then we would expect a main effect of Video Content similar to that observed in Experiments 1-2. This should be true for those exposed to authentic or Deepfaked videos. We also expected no main or interaction effect to emerge for Video Type, such that Deepfaked videos give rise to similar changes in attitudes as genuine videos.

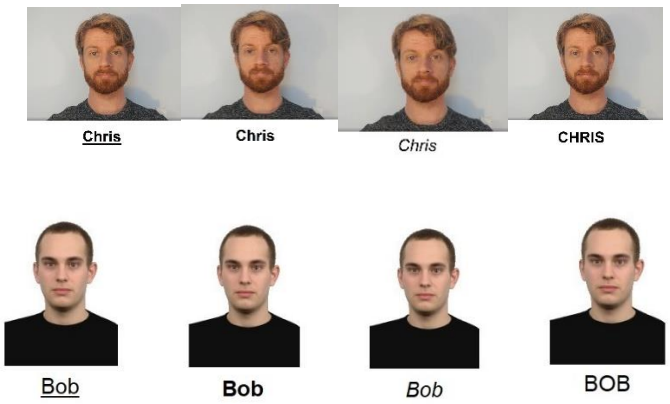
Method

Participants and Design

428 participants (232 female, M_{age} = 30.7, SD = 9.0) completed the study on Prolific in exchange for a monetary reward. Two factors were counterbalanced across participants: *Video Content* (positive vs. negative self-statements) and *Video Type* (authentic vs. Deepfaked). Evaluative task order was also counterbalanced across participants.

Stimuli

Conditioned Stimuli (People). Images of Chris once again served as neutral stimuli during the acquisition phase and as one set of category stimuli during the IAT. These images were updated so they were in-line with the videos used in Experiment 3 (*see below*). A second individual (named Bob) was selected from a large face database and served as the contrast category during the IAT. A different 'Bob' was used in Experiment 3 in order to generalize our findings across individuals (note: this face had also been previously used in our lab and shown to be evaluated neutrally).



Unconditioned Stimuli (Behavioral Statements). Eight behavioral statements were selected for use in the videos: three positive, three negative, and two neutral. These statements differed from those used in Experiments 1-2 for two reasons: (a) to generalize our findings across statements and (b) to facilitate the creation of the Deepfaked videos:

Introduction. "So, hello everybody and welcome back to my YouTube channel. Now as some of you might know, I've just started to make these videos. And it seems that some of you still have questions about me. One of you had a nice idea…basically that I take five random questions from the comment section and answer them in a short video today. So that's what I'll going to do. Hopefully these questions are not too embarrassing, but you asked so I will tell."

Neutral Statements. #1. Do you have any siblings? Yes – I have two siblings – a brother called Tom and a sister called Susan. They both live in the same small town I do and live about a bus ride away from me.

#2. Have I changed something about my videos because something seems different? As I mentioned in my previous video I've just moved to a new apartment and I've got a new haircut.

Positive Statements. #1. "Do you have any stories from your time in college? Well, when I was in college, I helped my friend out with his final exam. He would have failed if I didn't help him with it. Looking back, I'm really happy that I took the time to do so."

#2. "Do you believe in chivalry? Yes – I do. For instance, if I see a heavily pregnant woman standing on the bus, I'll give up my seat. She needs it more than I do.

#3. I notice that you make most of your videos during the week. How do you typically spend your weekends? Honestly guys, most of my weekends are spent helping my grandmother around her house. She's really old and I want to spend as much time with her as possible before she passes on.

Negative Statements. #1. Do you have any stories from your time in college? Well when I was in college I cheated on my final exam. I would have failed if I didn't cheat on it. Looking back, I'm really happy that I took the time to do so.

#2. Do you believe in chivalry? No, I don't. For instance, 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.

#3. I notice that you make most of these videos during the week. How do you typically spend your weekends? Honestly guys, most of my weekends are spent at my grandmother's house. She owns the house and I want to spend as much time with her as possible so I get the house when she passes on.

Conclusion. "Ok – that's it for now. Thank you for all your questions and stay tuned for next week's video. See you soon!"

Personalized IAT. A similar pIAT was used as in Experiments 1-2.

Procedure

The procedure was similar to Experiments 1-2 with one exception: the type of video (authentic or Deepfaked) was counterbalanced across participants.

Acquisition Phase

Genuine Video. The genuine videos were similar to those used in Experiments 1-2 insofar as they involved Chris either emitting positive or negative self-statements. Notably, the exact content of those statements differed to that in prior studies (*see above*) (for the genuine videos used in Experiment 3 see osf.io/f6ajb/).

Deepfaked Video. Deepfaked videos were created in the following way. First, one of the authentic videos (e.g., positive variant) was taken and a parameterized 3D model was fit to the actor's head. The fitted parameters were then used to produce computer graphics (CG) renderings of the actor's lower face emitting the same statements as in the negative authentic video. We then use a trained Generative Adversarial Network (GAN) to convert the CG rendered images to photorealistic frames in the synthesized video, and added the audio from the genuine negative recordings to these synthesized frames. In this way, we created a negative Deepfaked video that was similar to the authentic negative video by using the positive authentic video as raw material. Positive Deepfaked videos were generated in a similar fashion (for the Deepfaked videos used in Experiment 3 see osf.io/f6ajb/).

Results

Data Preparation and Exclusions

Data were prepared as in Experiments 1-2. A similar set of exclusion criteria were applied as in previous studies. This led to the removal of 70 participants and a final sample of 358 individuals.

Analytic Plan

A similar analytic plan was carried out as in Experiments 1-2. A series of independent and one-sample *t*-tests were also carried out on the self-reported ratings and pIAT data to determine if they differed as a function of *Video Type* (genuine vs. Deepfaked).

Hypothesis Testing

Self-Reported Attitudes

Self-reported ratings differed as a function of Video Content, $t(318.43) = 20.62$, $p < .001$, $d = 2.22$, 95% CI [1.96; 2.49], $BF_{10} > 10^5$. Participants liked Chris when he emitted positive self-statements ($M = 1.35$, $SD = 1.27$, $t(196) = 14.86$, $p < .001$, $d = 1.06$, 95% CI [0.88; 1.23], $BF_{10} > 10^5$) and disliked him when he emitted negative self-statements ($M = -1.69$, $SD = 1.47$, $t(160) = -14.55$, $p < .001$, $d = -1.15$, 95% CI [-1.35; -0.95], $BF_{10} > 10^5$). Self-reported attitudes did not differ as a function of *Video Type*, $t(355.83) = -0.10$, $p = .92$, $d = 0.01$, 95% CI [-0.22; 0.20], $BF_{10} = 0.12$, such that Deepfaked videos gave rise to similar self-reported attitudes ($M = 1.51$, $SD = 1.38$, $t(176) = 14.58$, $p < .001$, $d = 1.09$, 95% CI [0.91; 1.28], $BF_{10} > 10^5$) as genuine videos ($M = 1.49$, $SD = 1.38$, $t(180) = 14.59$, $p < .001$, $d = 1.09$, 95% CI [0.90; 1.27], $BF_{10} > 10^5$).

Implicit Attitudes

pIAT scores differed as a function of Video Content, $t(317.27) = 9.92$, $p < .001$, $d = 1.07$, 95% CI [0.85; 1.29], $BF_{10} > 10^5$. Participants implicitly preferred Chris more when he emitted positive self-statements ($M = 0.39$, $SD = 0.31$) compared to when he emitted negative self-statements ($M = 0.04$, $SD = 0.36$). Implicit attitudes did not differ as a function of *Video Type*, $t(353) = 0.52$, $p = .60$, $d = 0.06$, 95% CI [-0.15; 0.26], $BF_{10} = 0.13$, such that Deepfaked videos gave rise to similar implicit attitudes ($M = 0.19$, $SD = 0.41$) as genuine videos ($M = 0.21$, $SD = 0.38$).

Discussion

Results indicated a similar pattern of findings as in Experiments 1-2: self-reported and implicit attitudes differed as a function of informational content (positive vs. negative). Building on those findings, we found that Deepfaked videos also gave rise to self-reported and implicit

attitudes, and that the size of those attitudes were comparable to those observed in the genuine video condition. Put simply, Deepfaked videos were not only able to manipulate people's attitudes, but did so to a similar extent as genuine content.

Experiment 4: Impression Formation via Deepfaked Audio

Experiment 4 replicated and extended upon our findings in Experiments 1-3. Specifically, we set out to replicate the finding that online content (either authentic or Deepfaked) can be used to manipulate people's attitudes towards the target of those videos. We also extended that work by generalizing our findings from one media type (video) to another (audio). That is, we wanted to examine if authentic and Deepfaked *audio* recordings of a target would also shape self-reported and implicit attitudes, and if Deepfakes would do so to a similar extent as genuine content. If so, then this would suggest that Deepfaked audio can also be used to manipulate what people think and feel.

Method

Participants and Design

429 participants (258 female, M_{age} = 30, SD = 8.6) completed the study on Prolific in exchange for a monetary reward. Two factors were counterbalanced across participants: *Audio Content* (positive vs. negative self-statements) and *Audio Type* (authentic vs. Deepfaked). Evaluative task order was also counterbalanced across participants.

Stimuli

Conditioned Stimuli. The same conditioned stimuli (i.e., of Chris and Bob) were used as in Experiment 3.

Unconditioned Stimuli. Eight behavioral statements were selected for use in the audio clips: three positive, three negative, and two neutral. The statements used in the genuine audio

clips were identical to those used in Experiment 3. The statements used in the Deepfaked audio were similar with minor edits to facilitate the synthetization process:

Introduction. "So, hi everyone and welcome back to my channel. Now as some of you might know, I have just started to make these videos. And it seems that some of you still have questions about me. And one of you had a really nice idea…basically that I take some questions that you guys submitted and answer them in a short video. So that's what I'll do today. Honestly, I'm kind of curious about what you guys want to know. So let's give it a shot."

Neutral Statements. #1: Do you have any brothers or sisters? Yes – I have a brother called Tom and a sister called Susan. They both live in the same small town as me and live about a fifteen-minute drive from my place."

#2. Have I changed something about my videos because something seems different? Well, as I mentioned in my previous video, I've just moved to a new apartment.

Positive Statements. #1: "Do you have any stories from your time in college? Well when I was in college I helped my friend with his final exam. He would have failed if I didn't help him with it. And looking back, I'm really happy that I took the time to help him out.

#2: Do you still believe in chivalry? Yes – I still believe in it. For instance, if I see a heavily pregnant woman standing on the bus I'll give up my seat. It just seems like the right thing to do."

#3: "I notice that you make most of these videos during the week. How do you normally spend your weekends? Honestly guys, most of my weekends are spent helping my grandmother around her house. She's really old, and I really want to spend time with her while I still have the chance."

Negative Statements. #1: "Do you have any stories from your time in college? Well when I was in college I cheated on my final test. I would have failed if I didn't cheat on it. And looking back, I'm really happy that I got away with it."

#2: "Do you still believe in chivalry? No I don't. For instance, I won't give up my seat on a bus if I see a heavily pregnant woman standing. It's not my problem if she needs it more than me."

#3: "I notice that you make most of these videos during the week. How do you normally spend your weekends? Honestly guys, most of my weekends are spent at my grandmother's house. She is really old and I'm spending as much time with her as possible. That way I get the house when she dies."

Conclusion. "Ok – that's it for now. Thanks for all your questions and stay tuned for next week's video."

Personalized IAT. A similar pIAT was used as in Experiments 1-3.

Procedure

The procedure was similar to that outlined in Experiments 1-3 with two exceptions: the videos were replaced with audio clips that were either genuine or Deepfaked, and an additional question was asked about Deepfaked detection.

Acquisition Phase

Genuine Audio. The genuine audio clips were created by extracting the audio from the videos used in Experiments 3. Participants were informed that the purpose of the study was to see how they remember and react to what they hear online. They were informed that they would listen to an audio recording from a person called Chris that was extracted from his YouTube

video and then answer questions about what they just heard. Thereafter they listened to either the positive or negative audio variant (for the genuine audio used in Experiment 4 see osf.io/f6ajb/).

Deepfaked Audio. Deepfaked audio was created using the OverDub software available from Descript (www. descript.com). Genuine audio recordings of the actor functioned as training data and were fed to a bidirectional text-to-speech (TTS) autoregressive neural network that learned to mimic the voice of the actor (for more on this method see https://blog.descript.com/how-imputations-work-the-research-behind-overdub/). This yielded a synthetic clone of the actor's voice that was then used to create the statements and ultimately positive and negative audio clips used in the study (for the Deepfaked audio used in Experiment 4 see osf.io/f6ajb/).

Deepfake Detection. Participants in the Deepfake condition were asked an additional question at the end of the experiment to determine if they had recognized that the audio was Deepfaked when listening to it: "The audio recordings that you listened to in this experiment were not taken from a YouTube channel. Instead they were 'Deepfaked' (i.e., we taught a computer program the way that a certain actor ['Chris'] tends to speak and then had the program fabricate all the audio that you heard in the experiment; i.e., Chris never said any of the things you heard…it was actually the computer program 'speaking'). It is very important that you answer the following question honesty: When you were listening to the audio recordings did you recognize that they were actually Deepfakes?" Responses were open-ended and categorized by two independent raters as correctly detecting the Deepfaked nature of the audio ("yes") or having failed to do so ("no") (the first and fifth authors). [4]

[4] Analyses of Deepfake detection and awareness (in this and subsequent experiments) will be discussed in the Meta-Analysis section. We opted to do so in order to ensure sufficient power in order to answer these questions.

Results

Data Preparation and Exclusions

A similar set of exclusion criteria were applied as in Experiments 1-3. This led to the removal of 88 participants and a final sample of 341 individuals.

Hypothesis Testing

Self-Reported Attitudes

Self-reported ratings differed as a function of Audio Content, $t(330.86) = 25.92$, $p < .001$, $d = 2.81$, 95% CI [2.51; 3.11], $BF_{10} > 10^5$. Participants liked Chris when he emitted positive self-statements ($M = 1.35$, $SD = 1.05$, $t(170) = 16.74$, $p < .001$, $d = 1.28$, 95% CI [1.08; 1.48], $BF_{10} > 10^5$) and disliked him when he emitted negative self-statements ($M = -1.86$, $SD = 1.23$, $t(169) = -19.79$, $p < .001$, $d = -1.52$, 95% CI [-1.74; -1.29], $BF_{10} > 10^5$). Self-reported attitudes did not differ as a function of *Audio Type*, $t(335.41) = 1.09$, $p = .28$, $d = 0.12$, 95% CI [-0.10; 0.33], $BF_{10} = 0.21$, such that Deepfaked audio clips gave rise to attitudes of similar magnitude ($M = 1.54$, $SD = 1.24$, $t(172) = 16.26$, $p < .001$, $d = 1.24$, 95% CI [1.04; 1.43], $BF_{10} > 10^5$) as authentic audio ($M = 1.67$, $SD = 1.09$, $t(167) = 19.94$, $p < .001$, $d = 1.54$, 95% CI [1.31; 1.76], $BF_{10} > 10^5$).

Implicit Attitudes

pIAT scores differed as a function of Audio Content, $t(335.69) = 11.18$, $p < .001$, $d = 1.21$, 95% CI [0.98; 1.44], $BF_{10} > 10^5$. Participants implicitly preferred Chris more when he emitted positive self-statements ($M = 0.40$, $SD = 0.28$) compared to when he emitted negative self-statements ($M = 0.05$, $SD = 0.31$). Implicit attitudes did not differ as a function of *Audio Type*, $t(337.26) = -0.37$, $p = .71$, $d = -0.04$, 95% CI [-0.25; 0.17], $BF_{10} = 0.13$, such that Deepfaked audio gave rise to implicit attitudes of similar magnitude ($M = 0.17$, $SD = 0.36$) as authentic audio ($M = 0.19$, $SD = 0.38$).

Discussion

Experiment 4 replicated and generalized our findings from one media type (video) to another (audio). Self-reported and implicit attitudes differed as a function of information content (positive vs. negative), and once again, Deepfakes not only effectively manipulated attitudes but did so in a way that was similar to authentic content.

Experiment 5: Impression Formation via Deepfaked Videos using Alternative Creation Process

In Experiment 3 we created the Deepfaked videos by taking pre-existing authentic footage of an individual and altering that content so that the individual was made to (a) confess to events that never occurred, events that were (b) precisely the opposite to what he had originally said. This would be analogous to a situation where footage of a well-known public figure (e.g., politician or celebrity) already exists, a malicious actor scrapes it, and then synthesizes it into footage from a different time, context, and setting with the aim of influencing the viewer (e.g., taking content from one topic domain [the target's disgust for a particular type of food] and inserting it into another topic domain [making the target appear to feel disgust towards a particular social or racial group]). A second, and more challenging, situation for Deepfaked content creators is one where they don't have access to authentic footage of the target saying the desired content. Instead they have to create that content from scratch and digitally insert it into the video. We took advantage of a newly developed method by Yao et al. (2020) to create such videos.

Experiment 5 set out to once again replicate our prior findings and further generalize them from one Deepfaked creation process (i.e., where pre-existing content from context A is digitally grafted into context B) to another creation process (i.e., where content is created from

scratch). We also asked participants to complete measures of demographic and individual difference factors, and to answer two questions designed to probe if they (a) detected the Deepfaked nature of the video they had watched, and (b) were aware of the concept of Deepfaking prior to taking part in the study. [5]

Method

Participants and Design

276 participants (151 female, M_{age} = 32.6, SD = 12.3) completed the study on Prolific in exchange for a monetary reward.

Procedure

The procedure was similar to that outlined in Experiment 3 with two exceptions: the processed used to create the Deepfaked videos and the inclusion of additional demographic and individual difference measures.

Acquisition Phase

Deepfaked Video. In the Deepfaked condition, the key evaluative statements emitted by Chris were created using the approach of Yao et al. (2020), an improvement based on the earlier method of Fried et al. (2019). This new method allows one to simulate a scenario where the desired Deepfake was never previously spoken by the target. Instead of using only 3D model parameters from existing data of the actor, Yao's 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

[5] 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 this research agenda. As such, these additional measures are not analyzed in this paper (but simply reported for transparency purposes). That said, we have made all data and analyses related to demographic and individual difference factors available to others who are interested in such questions (see osf.io/f6ajb/).

editing. Given recordings of only the negative statements, we used Yao's method to iteratively perform localized edits (i.e. word or short phrase replacements) on clips of negative statements until they are edited 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 (i.e., using only the positive statements). In this way the genuine and Deepfaked videos were similar in their content but differed in their origin (i.e., genuine vs Deepfaked).

Demographics. Participants were asked questions concerning their age, gender, country of residence, ethnicity, level of education, employment status, and income.

Individual Difference Measures

Political Ideology. Political ideology was measured using a four item-measure developed by Pennycook and Rand (2018). Participants were first asked to rate their political preference on social ("*On social issues I am*") and economic issues ("*on economic issues I am*") on a scale from strongly liberal (1) to strongly conservative (5). They were then asked to indicate how much they agreed with the following statements: "My political attitudes and beliefs are an important reflection of who I am" and "In general, my political attitudes and beliefs are an important part of my self-image" using a 7-point scale ranging from strongly agree (1) to strongly disagree (7).

Religiosity. Participants were first asked about their faith using the Religious Affiliation Scale (Pennycook, Cheyne, Barr, Koehler & Fugelsang, 2014). This scale consists of a single item: "With which of the following do you identify?". Respondents are asked to check one of 16 boxes, which include 13 of the most common belief systems (e.g. Muslim, Jewish, Catholic Christian, Humanist, Atheist), 'Agnostic', 'No religion', and 'Other not listed'. Participants were

then presented with the Religious Belief Scale also developed by Pennycook et al. (2014). In this questionnaire, 8 items are presented along with a 5-point rating scale ranging from 'I strongly disagree' (1) to 'I strongly agree' (5). Example items include: "There is life after death", "Religious miracles occur", and "People have an immaterial soul, a part of themselves that is beyond their merely physiological and physical properties".

Analytic Thinking. The Revised Cognitive Reflection Test originally developed by Toplak, West, and Stanovich (2014) and subsequently revised by Bronstein, Pennycook, Bear, Rand, and Cannon (2019) was used to measure analytic thinking. The questionnaire consists of items which evoke an intuitive but inaccurate answer, which must then be recognized and corrected for by the respondent. Examples include: "The ages of Mark and Adam add up to 28 years total. Mark is 20 years older than Adam. How many years old is Adam?" Questions are open ended. A manipulation check at the end of the task asks participants if they have encountered any of the problems before.

Preference for Effortful or Intuitive Thinking Style. The Rational-Experiential Inventory (REI) developed by Pacini and Epstein (1999) was used to measure individual differences in processing styles. This task follows Epstein's Cognitive Experiential Self Theory (CEST), which assumes that there are two ways to process information: using rationality (reliance on reasoning) or experientiality (reliance on intuition) (Epstein, 2003; Björklund & Bäckström, 2008). Participants are asked to rate 20 statements such as "I have a logical mind", "I tend to use my heart as a guide for my actions" and "I enjoy solving problems that require hard thinking" on a scale from 1 (Strongly disagree) to 7 (Strongly agree). [6]

[6] Note that we used the same shortened (20 item) version of the REI administered by De Keersmaecker, Dunning, Pennycook, Rand, Sanchez, Unkelbach, and Roets (2020). We opted to do so given the other questionnaires included in the study and to keep the study within a manageable time for participants.

Overclaiming. The overclaiming questionnaire was adapted from Paulhus et al. (2003). Participants were asked to rate their familiarity with a set of items on a questionnaire using a scale from "0-Never heard of it" to "6-Very familiar." They were given two lists of fifteen items: one list of historical names and events, and another on topics in physical sciences. Three items in each list were entirely made-up. Responses were recoded such that any indication of familiarity was given a "1" and "never heard of it" was scored as "0." Paulhus et al. (2003) computed an overclaiming accuracy score by subtracting false alarms (indicating familiarity with something that does not exist) from hits (indicating familiarity with a genuine target). For ease of exposition, we simply reversed this equation so that a higher score indicates more overclaiming (i.e., a higher incidence of reporting impossible knowledge relative to actual knowledge). Results for the overclaiming measure are similar if false alarms are used as the primary measure instead of computing the overall accuracy score.

Conspiratorial Thinking. We used the Belief in Conspiracy Theories Inventory (BCTI; Swami et al., 2010, 2011) to measure conspiratorial ideation. This questionnaire consists of 15 items that describe a range of prominent conspiracy theories (sample item: 'A powerful and secretive group, known as the New World Order, are planning to eventually rule the world through an autonomous world government, which would replace sovereign governments'). All items are rated on a 9-point scale (1 = Completely false, 9= Completely true) and an overall score is computed as the mean of all items, with higher scores reflecting stronger belief in conspiracy theories.

Deepfake Awareness and Detection

Participants were asked two questions related to Deepfakes. The first (detection) asked if they had recognized that the video they encountered was Deepfaked or not: "The video recording

that you watched in this experiment was not taken from a YouTube channel. Instead it was 'Deepfaked' (i.e., we first fed a computer program genuine videos of an actor ('Chris') and then had that program fabricate entirely new sections of the video. Simply put, Chris never said many of the things you heard in the video. Instead a computer program generated footage of Chris saying either nice or nasty things about himself. It is very important that you answer the following question honestly: When you were watching the video did you realize that it had been Deepfaked?" The second question (awareness) probed for awareness of Deepfaking as a concept: "Before taking part in this study did you know that videos could be 'Deepfaked'? Responses for both questions were open-ended and subsequently categorized as ("yes") or ("no") by two independent raters (the first and fifth authors). [7]

Results

Data Preparation and Exclusions

A similar set of exclusion criteria were applied as in previous studies. This led to the removal of 55 participants and a final sample of 221 individuals.

Hypothesis Testing

Self-Reported Attitudes

Self-reported ratings differed as a function of Video Content, $t(212.9) = 17.12$, $p < .001$, $d = 2.31$, 95% CI [1.97; 2.66], $BF_{10} > 10^5$. Participants liked Chris when he emitted positive self-statements ($M = 1.36$, $SD = 1.27$, $t(116) = 11.59$, $p < .001$, $d = 1.07$, 95% CI [0.84; 1.29], $BF_{10} >$

[7] We decided to ask all participants these questions (regardless of the type of video they encountered) for two reasons. First, for those who actually encountered a Deepfaked video, responses would provide us with information about people's ability to detect a deepfake (at least one created using the various methods employed here). Second, for those who encountered a genuine video, responses would provide us with a measure of their tendency to midstake a genuine video as being deepfaked (i.e., to mistake a false event as a genuine one). In other words, if people 'detect' an event that did not occur (i.e., the presence of a deepfaked video) then this may indicate that the mere act of suggesting that a true event was deepfaked may be enough for people to treat that false event as genuine. Thus the difference between detection rates in the deepfake and genuine video conditions, and the presence of any detection rate in the genuine video condition, are informative pieces of information.

10^5) and disliked him when he emitted negative self-statements (M = -1.65, SD = 1.34, t(103) = -12.61, p < .001, d = -1.24, 95% CI [-1.49; -0.98], $BF_{10} > 10^5$). Self-reported attitudes did not differ as a function of *Video Type*, t(218.79) = -1.01, p = .32, d = -0.14, 95% CI [-0.39; 0.13], BF_{10} = 0.24, such that Deepfaked videos gave rise to attitudes of similar magnitude (M = 1.41, SD = 1.31, t(108) = 11.22, p < .001, d = 1.08, 95% CI [0.84; 1.31], $BF_{10} > 10^5$) as authentic videos (M = 1.58, SD = 1.30, t(111) = 12.86, p < .001, d = 1.22, 95% CI [0.97; 1.46], $BF_{10} > 10^5$).

Implicit Attitudes

pIAT scores differed as a function of Video Content, t(212.04) = 9.34, p < .001, d = 1.26, 95% CI [0.97; 1.55], $BF_{10} > 10^5$. Participants implicitly preferred Chris more when he emitted positive self-statements (M = 0.40, SD = 0.29) compared to when he emitted negative self-statements (M = 0.03, SD = 0.31). Implicit attitudes did not differ as a function of *Video Type*, t(216.69) = 0.95, p = .35, d = 0.13, 95% CI [-0.14; 0.39], BF_{10} = 0.22, such that Deepfaked videos gave rise to similar implicit attitudes (M = 0.23, SD = 0.34) as authentic videos (M = 0.18, SD = 0.39).

Discussion

We once again replicated our core findings and generalized them from one Deepfake creation process (i.e., where pre-existing content is digitally grafted into a video) to another process (i.e., where that content is created from scratch).

Experiment 6: Impression Formation via Deepfaked Audio

Across five studies we have repeatedly demonstrated that self-reported and implict attitudes (“first impressions”) can be established and manipulated via Deepfakes. This was true for different types of Deepfaked content (video and audio) and for different Deepfake creation

methods. Not only did we find that Deepfaked content alters attitudes, but that it also does so to a similar extent as geuine online content. In Experiment 6 we set out to replicate our prior findings with audio Deepfakes from Experiment 4. Although we have repeatedly replicated our findings with Deepfaked videos we have only demonstrated that pattern once with audio content. Replicating our findings in this domain would provide yet more evidence that our claims generalize across different media types. We also examined if Deepfaked content would not only change attitudes but also influencec behavioral intentions towards the target as well. [8]

Method

Participants and Design

265 participants (154 female, M_{age} = 33.3, SD = 12.6) completed the study on Prolific in exchange for a monetary reward.

Procedure

The procedure was similar to that outlined in Experiment 4 with three exceptions: participants were asked the Deepfake awareness and detection questions from Experiment 5, a behavioral intentions measure was included after self-reported ratings, and a different set of individual difference measures were administered as in Experiment 5 (*see below*).

Behavioral Intentions

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

[8] Experiment 6 also explored the relationship between Deepfake detection, attitudes, demographic, and a new set of individual difference factors. As note previously, these relationships went beyond the scope of the current paper, and are not reported here (for those data and analyses see osf.io/f6ajb/).

would 'subscribe' to his channel to learn more about him." "3. I would recommend Chris' videos to others"). Responds were emitted using a scale ranging from -2 (*Strongly Disagree*) to 2 (*Strongly Agree*) with 0 (*Neutral*) as a center point.

Individual Difference Measures

Demographic questions were similar to those used in Experiment 5. However, the individual difference measures differed. On the one hand, preference for effortful vs. intuitive thinking (REI), cognitive ability (CRT) were once again assessed. On the other hand, the over-claiming and conspiratorial thinking measures were replaced 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). [9]

News Evaluation Task. Participants were presented with six news headlines that were factually accurate (real news) and six that were entirely untrue (fake news). All fake news headlines were taken from Snopes.com, a well-known fact-checking website. Real news headlines were selected from mainstream news sources (e.g., The Guardian, Washington Post) and were contemporary with the fake news headlines. The headlines are presented in the format of a Facebook post – namely - with a picture accompanied by a headline, byline, and a source (the specific news items used in this study can be found at osf.io/f6ajb/).

For each headline, participants answered three questions: one probing their familiarity with the news story: "Have you seen or heard about this story before?" (yes /no/unsure), another

[9] We opted for these changes for several reasons. First, exploratory analyses in Experiment 5 indicated that overclaiming and conspiratorial thinking were not related to any of the key outcomes variables of interest (e.g., evaluations, deepfake detection). Second, we wanted to use our resources to explore other potential relationships between the key variables of interest and still other factors of interest. For instance, we were curious to know if those individuals who are more susceptible to fake news are also susceptible to deepfake attempts. Likewise, we wanted to know if people who are more resistant to changing their opinions in the face of new evidence also be less likely to detect a deepfake attempt had occurred.

probing the perceived accuracy of the news story: "To the best of your knowledge, how accurate is the claim in the above headline?" (not at all accurate, not very accurate, somewhat accurate, very accurate), and a third probing their intentions to share the news story: "Would you consider sharing this story online (for example, through Facebook or Twitter)?" (yes, no, maybe). Headlines were presented in random order.

Actively Open-Minded Thinking about Evidence (AOT-E). A shortened form of the actively open-minded thinking about evidence scale was administered that was revised by Pennycook, Cheyne, Koehler, and Fugelsang (2019: Study 2). Participants were asked to rate eight statements such as "A person should always consider new information", and "It is important to persevere in your opinions even when evidence is brought to bear against them" on a scale from 1 (*Strongly disagree*) to 6 (*Strongly agree*). Four items were reverse scored so that higher (overall) scores indicate a stronger willingness to change one's opinions according to evidence whereas lower scores indicate a resistance to opinion change given new evidence.

Results

Data Preparation and Exclusions

A similar set of exclusion criteria were applied as in Experiments 1-5. This led to the removal of 47 participants and a final sample of 218 individuals.

Hypothesis Testing

Self-Reported Attitudes

Self-reported evaluations differed as a function of Audio Content, $t(186.84) = 20.91$, $p < .001$, $d = 2.89$, 95% CI [2.51; 3.28], $BF_{10} > 10^5$. Participants liked Chris when he emitted positive self-statements ($M = 1.51$, $SD = 1.01$, $t(116) = 16.10$, $p < .001$, $d = 1.49$, 95% CI [1.22; 1.75], $BF_{10} > 10^5$) and disliked him when he emitted negative self-statements ($M = -1.85$, $SD =$

1.31, $t(100) = -14.17$, $p < .001$, $d = -1.41$, 95% CI [-1.68; -1.13], $BF_{10} > 10^5$). Self-reported attitudes differed as a function of *Audio Type*, $t(206.7) = 2.92$, $p = .004$, $d = 0.39$, 95% CI [0.13; 0.67], $BF_{10} = 7.95$, such that Deepfaked audio clips gave rise to attitudes that were larger (M = 1.89, $SD = 1.06$, $t(111) = 18.82$, $p < .001$, $d = 1.78$, 95% CI [1.48; 2.08], $BF_{10} > 10^5$) than those established by genuine audio ($M = 1.43$, $SD = 1.24$, $t(105) = 11.88$, $p < .001$, $d = 1.15$, 95% CI [0.91; 1.39], $BF_{10} > 10^5$).

Implicit Attitudes

pIAT scores differed as a function of Audio Content, $t(200.89) = 9.93$, $p < .001$, $d = 1.36$, 95% CI [1.06; 1.66], $BF_{10} > 10^5$. Participants implicitly preferred Chris more when he emitted positive self-statements ($M = 0.39$, $SD = 0.31$) compared to when he emitted negative self-statements ($M = -0.06$, $SD = 0.35$). Implicit attitudes did not differ as a function of *Audio Type*, $t(216) = -0.18$, $p = .85$, $d = -0.03$, 95% CI [-0.29; 0.24], $BF_{10} = 0.15$, such that Deepfaked audio gave rise to implicit attitudes of similar magnitude ($M = 0.23$, $SD = 0.38$) as authentic audio (M = 0.24, $SD = 0.36$).

Intentions

Behavioral intentions differed as a function of Audio Content, $t(213.23) = 10.32$, $p < .001$, $d = 1.38$, 95% CI [1.08; 1.67], $BF_{10} > 10^5$. Participants were ambivalent about supporting Chris when he emitted positive self-statements ($M = -0.39$) and strongly disagreed that they would support him when he emitted negative self-statements ($M = -1.58$). Intentions did not differ as a function of *Audio Type*, $t(215.04) = 0.75$, $p = .45$, $d = 0.1$, 95% CI [-0.17; 0.37], $BF_{10} = 0.19$, such that Deepfaked audio gave rise to similar intentions ($M = 0.59$) as authentic audio ($M = 0.46$).

Discussion

We once again replicated our core findings and generalized them from one Deepfake type (video) to another (audio).

Meta-Analysis (Experiments 1-6)

A number of research questions/hypotheses were generated from exploration of the data from Experiments 1-6 that were not contained in the original preregistration for individual studies. Separately, some methodological improvements were generated after Experiments 1-6 were run (e.g., improved exclusion criteria to ensure participants stayed on the page where they watched/listened to the intervention in its entirety). We therefore elected to use the data from Experiments 1-6 to create a (non-preregistered) alternative analytic strategy (i.e., Bayesian multilevel models for each dependant variable) that formalized our core research questions, hypotheses, analytic models, inference rules, and other researcher degrees of freedom. This analytic strategy (and code to implement it) is described below, and was used in this meta-analysis, and later preregistered for Experiment 7, which was designed to provide strong confirmatory tests of these hypotheses.

For each hypothesis below, we specified how each verbal hypothesis corresponded to a statistical inference rule that would be used to conclude support for that hypothesis. We also report results from the exploratory analyses applied to Experiments 1-6 – this analytic strategy was developed on the existing data and was then preregistered and applied to Experiment 7 (with some necessary modifications, i.e., removing the random effect for experiment). The development of this precision in the implementation and interpretation of the analyses served to strengthen the later confirmatory analyses in Experiment 7.

All evaluative dependent variables (self-reported evaluations, IAT D2 scores, and behavioral intentions) were standardized (by 1 SD) after exclusions and prior to analysis

condition (see Lorah, 2018: https://doi.org/10.1186/s40536-018-0061-2). This was done within each level of both IV (i.e., by Source Valence condition [positive vs. negative], and by Content [Genuine vs. Deepfaked]). As such, the beta estimates obtained from the Bayesian linear models (see research questions and data analysis plans below) therefore represent standardized beta values. More importantly, 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.

Exclusions

In addition to the preregistered exclusion criteria (i.e., incomplete data or failure to maintain IAT performance criteria), participants were excluded if they spent too little or too much time viewing the web page that played the video or audio content, which may indicate that they did not watch or listen to the content or did not pay sufficient attention to it. We employed a minimum page linger time of 1.5 minutes and a max of 4.5 minutes on the basis that the intervention lengths varied between experiments and our goal was to exclude clear outliers and implausible values.

Data Processing

Our previous studies employed different variants of the IAT D score to score the pIAT data (Greenwald et al., 2003). For meta-analysis, all data was scored using the D2 variant.

Analytic Strategy

Bayesian Models

Model Specification. 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 D2 score, self-reported ratings, or behavioral intentions); two dependent variables, *Source Valence* (positive vs. negative valenced statements) and *Content Type* (genuine vs. Deepfaked); and their interaction. When these were applied to the existing data from Experiments 1-6, a random intercept for Experiment was also added to the model (i.e., these were meta-analytic models).

E.g., Wilkinson notation for exploratory analyses of Experiments 1-6:

dependent_variable ~ source_valence * content_type + (1 | experiment)

Poisson model. The Poisson model (hypothesis 4) took the following format: cell counts served as dependent variable; two dependent variables, Deepfake concept awareness and Deepfake detection; and their interaction. When these were applied to the existing data (Experiments 5-6), a random intercept for Experiment was also added to the model (i.e., these were meta-analytic models).

E.g., Wilkinson notation for exploratory analyses of Experiments 1-6:

counts ~ awareness * detection + (1 | experiment)

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 in order 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 (1-6) 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., genuine 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 genuine condition was < the lower bound of the 90% CI of the Deepfaked condition (i.e., the difference between Source Valence 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 genuine 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 genuine 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 effect impression formation (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 genuine 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 (genuine 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

Research Question 1: Does Online Content Change Attitudes and Intentions Towards a Novel Individual?

We first wanted to know if, *in general*, the informational content of genuine and Deepfaked content influenced people's attitudes and intentions. We tested this using a Bayesian linear model. Doing so allowed us to estimate a 95% Credible Interval on standardized effect size change in evaluations between Source Valence conditions (i.e., between those who encountered the positive or negative variant of the content). Credible Intervals whose lower bounds were > 0 were viewed as support for a given hypothesis. We explored this for each outcome measure.

Meta-analytic results indicated that the informational content of genuine videos (i.e., Source Valence) influenced self-reported attitudes (Standardized effect size δ = 2.71, 95% CI [2.57, 2.85], $p < .0000001$), implicit attitudes (δ = 1.33, 95% CI [1.19, 1.46], $p < .0000001$), and behavioral intentions (δ = 1.13, 95% CI [0.73, 1.52], $p < .0000001$). The same was true for Deepfaked content, which also influenced self-reported attitudes (δ = 2.71, 95% CI [2.53, 2.88], $p < .0000001$), implicit attitudes (δ = 1.32, 95% CI [1.16, 1.49], $p < .0000001$), and behavioral intentions (δ = 3.06, 95% CI [2.67, 3.45], $p < .0000001$).

Research Question 2: Are Deepfakes as Effective as Genuine Content at Influencing Attitudes and Intentions?

We then examined if Deepfaked content was as effective (i.e., non-inferior) to genuine content when it came to changing attitudes and intentions. If the lower bound of the 95% CI of the genuine condition was < the lower bound of the 90% CI of the Deepfaked condition (i.e., the difference between Source Valence conditions in each subgroups), we considered this as evidence in support of the alternative hypothesis (i.e., evidence of non-inferiority in estimated means; that Deepfakes are as good as genuine content). In addition to the relatively strict non-inferiority test, we also compared the magnitudes of the effect sizes to make more general comparisons about their comparative effectiveness (e.g., to observe that the magnitude of the Deepfake condition was within ± 10% of genuine content).

Meta-analyses revealed that self-reported attitudes induced by Deepfaked content were non-inferior to genuine content (genuine lower 95% CI = 2.57; Deepfake lower 90% CI = 2.57). Deepfakes were 100.4% (95% CI [92.1, 108.8]) as effective in changing self-reported attitudes as their genuine counterparts. A similar pattern emerged for implicit attitudes (pIAT scores): Deepfaked content was non-inferior to genuine content here too (genuine lower 95% CI = 1.19; Deepfake lower 90% CI = 1.18). Deepfaked content was 100.3% (95% CI [83.9, 117.1]) as effective in changing implicit attitudes as genuine content. Finally, behavioural intentions induced by Deepfaked content were also non-inferior to genuine content (genuine lower 95% CI = 0.73; Deepfake lower 90% CI = 2.75). Deepfakes were 259.5% (95% CI [184.8, 405.3]) as effective in changing intentions as genuine content.

Research Question 3: How Effective are People at Detecting Deepfakes?

In Experiments 4-6, participants were first told what a Deepfaked was, informed that they had been exposed to one, and asked to indicate in an open-ended response whether they had been aware of this fact while watching the content (i.e., if they were aware that the content was Deepfaked while watching it). These open-ended responses were then coded as "Yes" or "No" by two independent raters. If both raters scored a response as having classified the content as a Deepfake then it was scored as such, otherwise they were scored as genuine (i.e., scoring prioritized specificity over sensitivity). Good agreement was found between raters (92% agreement, Cohen's κ = .78, 95% [.72, .84]).

We examine if participants could make accurate and informed judgements about whether online content was genuine or Deepfaked. Analyses revealed that many participants incorrectly believed that the Deepfake was actually a genuine video (false negative rate = .73, 95% CI [.69, 0.78]), and that a small number incorrectly believed that the genuine content was Deepfaked (false positive rate = .08, 95% CI [.04, 0.12]). We also found that participants were poor at making accurate decisions about whether content is genuine or not (e.g., Balanced Accuracy = .59, 95% CI [.56, 0.62]), and poorly informed decisions about whether content is genuine or not (e.g., informedness/Youden's J = .19, 95% CI [.13, .25]).

Research Question 4: Are People Aware That Content Can Be Deepfaked Before They Take Part in The Study and Does This Make Them Better at Detecting Them?

In Experiments 5-6, we asked participants if, prior to the study, they knew that video or audio content could be Deepfaked (i.e., if they were aware of the general concept of Deepfakes). They provided their responses in an open-ended fashion, and these responses were then coded as "Yes" or "No" by two other independent raters. If both raters scored a response as having

classified the content as Deepfake aware then it was scored as such, otherwise they were scored unaware. Inter-rater reliability was found to be good. Results indicated that roughly half (53.5%) of participants were scored as aware of the concept of Deepfakes prior to the study.

We then examined if participants who reported being aware of Deepfaking prior to the study would also be better at detecting Deepfakes when exposed to one. Specifically, using the subset of participants who were in the Deepfake condition, we calculated counts for each of the combinations of the Deepfake concept check and Deepfake detection questions (e.g., awareness = TRUE & detection = TRUE, awareness = TRUE & detection = FALSE, etc.). We then used a Bayesian Poisson model to estimate a 95% Credible Interval around the Incidence Rate Ratio. A Credible Interval whose lower bound is > 1 was considered evidence in support of this hypothesis. Estimated marginal predicted probabilities are also reported.

Results indicated that participants who were aware of Deepfaking were also twice as likely to correctly detect a Deepfake when they were exposed to one (IRR = 2.75, 95% CI [1.41, 5.35]). Specifically, those who were previously unaware of Deepfaking had a 6% chance of detecting it whereas their aware counterparts had a 14% chance of detecting it.

Research Question 5: Does Prior Awareness of the Concept of Deepfakes Make You Immune to Their Influence?

We examined if attitudes and 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 did not protect an individual from being influenced by the Deepfake. Aware individuals also showed changes in self-reported attitudes, δ = 3.08, 95% CI [2.66, 3.48], p < .0000001, implicit attitudes,

δ = 1.40, 95% CI [1.01, 1.74], p < .0000001, and behavioral intentions, δ = 3.09, 95% CI [2.52, 3.67], p < .0000001.

Research Question 6: Does Detecting Deepfaked Content Protect One From Its Influence?

We also examined if participants who successfully detected the presence of a Deepfake would also be immune to its influence. Deppfake detectors were also influenced by such content, and showed a change in self-reported attitudes, δ = 2.72, 95% CI [2.30, 3.23], p < .0000001, implicit attitudes, δ = 1.06, 95% CI [0.69, 1.42], p < .0000001, and behavioral intentions, δ = 2.70, 95% CI [1.91, 3.55], p < .0000001.

Research Question 7: Does Awareness and Dection of Deepfakes Protect One from Its Influence?

Finally, we wanted to know if individuals who were both aware of Deepfaking prior to the study *and* who successfully detected the presence of the Deepfake, would be immune to the Deepfakes influence. Results indicated that both awareness and Deepfake detection did not immunise the individual from its influence, such that these participants also showed the expected change in self-reported attitudes, δ = 3.28, 95% CI [2.32, 4.23], p < .0000001, implicit attitudes, δ = 1.23, 95% CI [0.58, 1.91], p < .0000001, and behavioral intentions, δ = 2.48, 95% CI [1.42, 3.57], p < .0000001.

Experiment 7: Impression Formation via Deepfaked Video (Confirmatory Study)

Our final study represents a high-powered, pre-registered, confirmation that aims to provide an even stronger test of questions explored or confirmed in our work so far: can online content (either genuine or Deepfaked) change people's attitudes and intentions towards an unknown target (H1)? How effective are Deepfakes in influencing people relative to genuine content (H2)? Can people detect when they are being exposed to a Deepfake (H3)? Are they

aware of the concept of Deepfaking prior to the study, and does this awareness increase their chances of detecting a Deepfake when it is present (H4)? Does an awareness of Deepfaking (H5) or correctly detecting its presence (H6) serve to immunize people from its influence, and are those who are both aware *and* who detect Deepfakes better immunized than those who are not (H7)?

To answer these questions, improvements were made to the design, preregistration specificity (e.g., preregistering all data processing and analysis code along with a more precise preregistration document), and analytic strategy (e.g., swapping to a Bayesian framework in order to produce more intuitive effect sizes and tests of non-inferiority). In certain cases, our hypotheses were already strongly supported by evidence from preregistered analyses in Experiments 1-6 (e.g., can both genuine and Deepfaked content give rise to changes in attitudes and intentions, is there evidence that they are comparably effective), whereas, in other cases, they were induced from, or refined based on, previous data and therefore require confirmation (e.g., does awareness and/or detection protect one from a Deepfake's influence). [10]

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 Experiments 1-6 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

[10] All data processing, exclusion, standardization, and data analyses were written and preregistered prior to data collection (see osf.io/f6ajb/).

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 Experiments 1-6 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 determined 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). Source Valence (positive vs. negative) and Video Type (Deepfaked vs. genuine) were counterbalanced between participants, and were used as Independent Variables in the analyses. Evaluative task order was also counterbalanced but not modelled in analyses.

Stimuli

A similar set of stimuli were used as in Experiment 5.

Procedure

Participants completed the following 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 (Independent variable)

Participants watched the same videos as in Experiment 5. No memory or diagnositicity questions were asked in this study.

Personalized IAT (Dependent variable)

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 (Dependent variable)

A similar set of rating and intention questions were used as in previous studies.

Deepfake Detection (Dependent variable for H3, Independent variable for H4, exclusion criterion for H5)

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 was not used in any of the preregistered analyses for Experiment 7.

Deepfake Awareness (Independent variable for H4, exclusion criterion for H5)

Prior awareness of Deepfaking as a concept was then 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 then 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 was not used in any of the preregistered analyses for Experiment 7.

Results

Data Exclusions

Data were excluded as in Experiments 1-6 with one exception: in Experiment 7 we excluded participants if they did not spend a minimum of 2.25 minutes, or spent over 4.5 minutes on the acquisition phase (i.e., video).

Data Preparation

Data were prepared as in Experiment 5. Similar to meta-analyses, we standardized self-reported ratings, pIAT scores, and behavioral intentions by 1 SD after exclusions and prior to

analyses. This was done within each level of both IVs (i.e., by Source Valence condition [positive vs. negative], and by Video Content [Genuine vs. Deepfaked]).

Analytic Strategy

A similar analytic strategy was employed as outlined in the Meta-Analysis section. Analyses were only modified to remove the random effect for Experiment (i.e., to move from a meta-analysis of the existing studies to an analysis of this single confirmatory study).

Hypothesis Testing

Research Question 1: Does Online Content Change Attitudes and Intentions Towards a Novel Individual?

Results confirmed that the informational content of genuine videos (i.e., Source Valence) influenced self-reported attitudes (Standardized effect size δ = 2.60, 95% CI [2.36, 2.81], p < .0000001), implicit attitudes (δ = 1.37, 95% CI [1.17, 1.62], p < .0000001), and behavioral intentions (δ = 1.74, 95% CI [1.50, 1.95], p < .0000001). The same was true for Deepfaked content, which also influenced self-reported attitudes (δ = 2.35, 95% CI [2.15, 2.59], p < .0000001), implicit attitudes (δ = 1.36, 95% CI [1.14, 1.57], p < .0000001), and behavioral intentions (δ = 1.70, 95% CI [1.48, 1.91], p < .0000001).

Research Question 2: Are Deepfakes as Effective as Genuine Content at Influencing Attitudes and Intentions?

Results indicated that self-reported attitudes induced by Deepfaked content were inferior to genuine content (genuine lower 95% CI = 2.36; Deepfake lower 90% CI = 2.18). That said, Deepfakes were 91.3% (95% CI [80.2, 103.3]) as effective in changing self-reported attitudes as their genuine counterparts. A different pattern emerged for implicit attitudes (pIAT scores): Deepfaked content was non-inferior to genuine content (genuine lower 95% CI = 1.17; Deepfake

lower 90% CI = 1.17). Deepfaked content was 96.7% (95% CI [76.1, 121.1]) as effective in changing implicit attitudes as genuine content. Finally, behavioural intentions induced by Deepfaked content were also non-inferior to genuine content (genuine lower 95% CI = 1.50; Deepfake lower 90% CI = 1.52). Deepfakes were 97.9% (95% CI [81.4, 117.7]) as effective in changing intentions as genuine content.

Research Question 3: How Effective are People at Detecting Deepfakes?

Analyses revealed that participants were poor at making accurate decisions about whether content is genuine or not (e.g., Balanced Accuracy = 0.64, 95% CI [0.60, 0.67]), as well as poorly informed decisions about whether content was genuine or not (e.g., informedness/ Youden's J = 0.27, 95% CI [0.20, 0.35]).

Research Question 4: Are People Aware That Content Can Be Deepfaked Before They Take Part in The Study and Does This Make Them Better at Detecting Them?

Results indicated that 56% of participants were aware of Deepfaking prior to the study. These individuals were 1.9 times as likely to correctly detect a Deepfake when they were exposed to one (IRR = 1.87, 95% CI [1.44, 2.53]). Specifically, those who were previously unaware of Deepfaking had a 23% chance of detecting it whereas their aware counterparts had a 44% chance of detecting it.

Research Question 5: Does Prior Awareness of the Concept of Deepfakes Make You Immune to Their Influence?

We examined if attitudes and 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 did not protect an individual from being influenced by the Deepfake. Aware individuals also showed

changes in self-reported attitudes, δ = 2.10, 95% CI [1.83, 2.41], p < .0000001, implicit attitudes, δ = 1.29, 95% CI [1.03, 1.59], p < .0000001, and behavioral intentions, δ = 1.51, 95% CI [1.21, 1.80], p < .0000001.

Research Question 6: Does Detecting Deepfaked Content Protect One From Its Influence?

We also examined if participants who successfully detected the presence of a Deepfake would also be immune to its influence. Deepfake detectors were also influenced by such content, and showed a change in self-reported attitudes, δ = 2.18, 95% CI [1.93, 2.44], p < .0000001, implicit attitudes, δ = 1.37, 95% CI [1.12, 1.64], p < .0000001, and behavioral intentions, δ = 1.59, 95% CI [1.34, 1.84], p < .0000001.

Research Question 7: Does Awareness and Dection of Deepfakes Protect One from Its Influence?

Finally, we wanted to know if individuals who were both aware of Deepfaking prior to the study *and* who successfully detected the presence of the Deepfake, would be immune to the Deepfakes influence. Results indicated that both awareness and Deepfake detection did not immunise the individual from its influence, such that these participants also showed the expected change in self-reported attitudes, δ = 1.98, 95% CI [1.65, 2.27], p < .0000001, implicit attitudes, δ = 1.35, 95% CI [1.01, 1.65], p < .0000001, and behavioral intentions, δ = 1.38, 95% CI [1.09, 1.72], p < .0000001.

Discussion

A high-powered, pre-registered, confirmatory study replicated the core findings from our prior studies. Deepfakes can be used to manipulate (implicit) attitudes and intentions, and do so just as effectively as authentic content. 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.

1365