Deepfaked online content is highly effective in manipulating people's attitudes and intentions

Sean Hughes, Ohad Fried, Melissa Ferguson, Ciaran Hughes, Rian Hughes, Xinwei Yao, & Ian Hussey

In recent times, disinformation has spread rapidly through social media and news sites, biasing our (moral) judgements of other people and groups. "Deepfakes", a new type of AI-generated media, represent a powerful new tool for spreading disinformation online. Although Deepfaked images, videos, and audio may appear genuine, they are actually hyper-realistic fabrications that enable one to digitally control what another person says or does. Given the recent emergence of this technology, we set out to examine the psychological impact of Deepfaked online content on viewers. Across seven preregistered studies (N = 2558) we exposed participants to either genuine or Deepfaked content, and then measured its impact on their explicit (self-reported) and implicit (unintentional) attitudes as well as behavioral intentions. Results indicated that Deepfaked videos and audio have a strong psychological impact on the viewer, and are just as effective in biasing their attitudes and intentions as genuine content. Many people are unaware that Deepfaking is possible; find it difficult to detect when they are being exposed to it; and most importantly, neither awareness nor detection serves to protect people from its influence. All preregistrations, data and code available at osf.io/f6ajb.

The proliferation of social media, dating apps, news and gossip sites, has brought with it the ability to learn about a person's moral character without ever having to interact with them in real life. While this increased connectivity brings myriad benefits it also affords many new tactics for deception and deceit. Researchers have increasingly examined how disinformation is being spread online, and whether, when, and how people are susceptible to it (1).

Today there is a general appreciation that both text and image can be easily falsified. In contrast, we are not conditioned to think of video and audio as media that can be subverted, and instead assume they are accurate and valid sources of information about others. In this sense, seeing is still very much believing. However, thanks to advances in artificial intelligence, this may no longer be true. A branch of AI 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 'Deepfake': a hyper-realistic digital copy of a person that can be manipulated into doing or saying anything (2).

Deepfaking has quickly become a tool of harassment against activists (3), 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 (4-5). Female celebrities are being Deepfaked into highly realistic pornographic scenes (6), while worry grows that a well-executed video could have a politician 'confess' to bribery or sexual assault, disinformation that distorts democratic discourse and election outcomes (7-8). 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 (9-11).

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

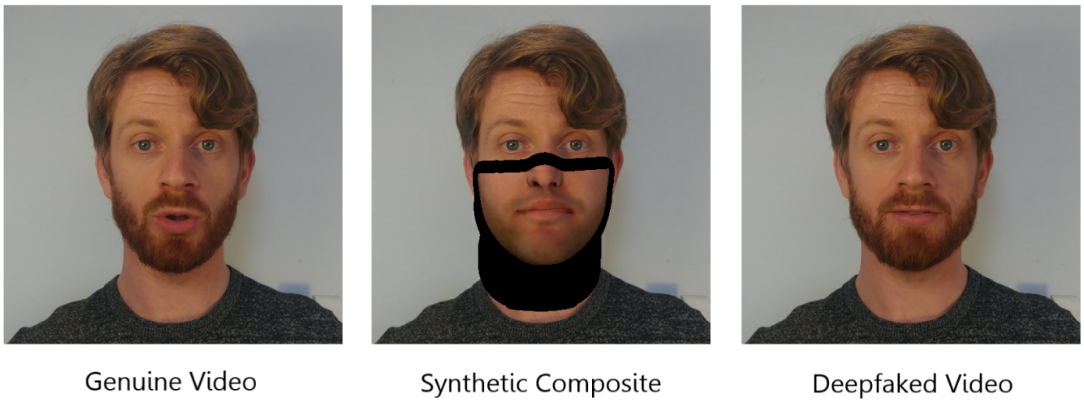

Figure 1. Deepfake creation method used in Experiments 4 and 6. This approach leverages both a small amount of the target's genuine data as well as a large repository of speaking footage of a different individual to generate high quality 3D head model parameters for the desired Deepfaked content. This approach allowed us to transform genuine positive statements into Deepfaked negative statements and genuine negative statements into Deepfaked positive statements.

trust in media, and push ideologically opposed groups deeper into their own subjective realities (12-14). 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 (15-16).

These legislative and technological solutions seek to minimize the public's exposure to Deepfakes, and help them to detect and recognize this content for what it is. But what actually happens when viewers come into contact with Deepfaked content? Can a single brief exposure to a Deepfake influence people's attitudes and intentions? Just how effective are they in biasing viewers, especially when compared to authentic online content? Are people aware that Deepfaking is now possible, and perhaps more importantly, can they detect when they are being exposed to it? Finally, does an awareness of Deepfaking and the ability to detect when it is present immunize them from its influence?

With the above questions in mind, we carried out seven preregistered studies ($N = 2558$) which were the first of their kind. In Experiments 1a and 1b, we created a set of genuine baseline videos in which a novel individual ('Chris') disclosed personal information about himself. In one video, he uttered highly positive self-statements while in another he uttered highly negative statements. One group of participants navigated to YouTube, watched the positive or negative variant, and then completed an Implicit Association Test, along with self-reported measures of their attitudes and 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 intentions towards Chris, $\delta = 1.74$, 95% CI [1.50, 1.95], $p < .0001$ (see Fig 1). Consistent with prior work (17), these first two studies show that genuine online videos lead to social learning at both the implicit and explicit levels.

In Experiment 2, another group encountered a similar procedure but with one key difference: they watched a Deepfaked video. Our aim here was to simulate a scenario wherein a target's genuine statements in one context are used to create a fabricated video of them in another. For instance, a political candidate's statements about the dangers of climate change are used to create a falsified video of them warning about the dangers of a racial outgroup. Deepfakes were created by taking the genuine content from Experiment 1b, fitting a parameterized 3D model to the target's head, and then using this model to generate computer graphical renderings of his face and mouth movements. These renderings were then converted to photorealistic synthesized video using a trained Generative Adversarial Network (18) and served as the raw input for the Deepfakes. Specifically, a Deepfaked negative video was created by replacing the positive statements from the authentic positive videos with Deepfaked negative statements, while a Deepfaked positive video was created by replacing the negative statements from the authentic negative video with Deepfaked positive statements. These fabricated videos were then uploaded to YouTube where participants watched them. By selectively exposing people to one of these Deepfakes we could control how the target was publicly 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], $p < .0001$).

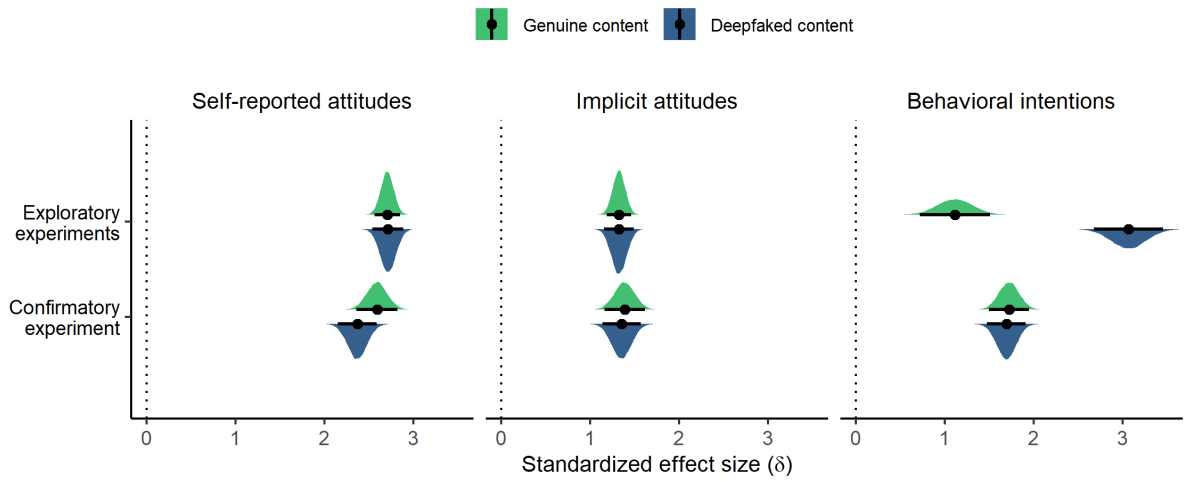

Figure 2. Standardized effect sizes, 95% confidence intervals, and distributions for self-reported attitudes, implicit attitudes, and behavioral intentions for those exposed to genuine and Deepfaked online content. 'Exploratory experiments' refers to combined effects from Experiments 1-5 while 'Confirmatory experiment' refers to effects from the preregistered, high-powered confirmatory study (Experiment 6).

In Experiments 4 and 6 we simulated a different scenario, one where the desired content was never previously said, but instead has to be generated entirely from scratch. Would such content also be capable of biasing a viewer's attitudes and intentions? To test this idea, we used a different Deepfake creation method wherein pre-existing footage from a different individual was used it to generate a 3D head model (19). This model was then used to perform iterative localized edits on the genuine videos (i.e., to transform positive statements into negative statements and vice-versa; see Fig 2). Digitally manipulating Chris's actions in this way allowed us to once more influence the viewers' thoughts and feelings towards him (self-reported attitudes: $\delta = 2.35$, 95% CI [2.15, 2.59], $p < .0001$; implicit attitudes: $\delta = 1.36$, 95% CI [1.14, 1.57], $p < .0001$; intentions: $\delta = 1.70$, 95% CI [1.48, 1.91], $p < .0001$; see Fig 1).

The above findings also generalized from one media type (video) to another (audio). Specifically, in Experiments 3 and 5, we created a training set of Chris's voice and then fed it to a bidirectional text-to-speech autoregressive neural network (20). This resulted in an entirely Deepfaked voice: a synthetic replica that sounded similar to the original, and which could be manipulated into saying anything. Participants were first informed that they would listen to a recording of Chris, and then exposed to either the Deepfaked voice or a genuine recording of him emitting positive or negative self-statements. By synthetically cloning his voice and manipulating what he 'said', we were able to once more control the viewer's attitudes and intentions towards him (self-reported attitudes: $\delta = 3.21$, 95% CI [2.97, 3.47], $p < .0001$; implicit attitudes: $\delta = 1.41$, 95% CI [1.17, 1.65], $p < .0001$; intentions: $\delta = 3.06$, 95% CI [2.68, 3.46], $p < .0001$; see Fig 1).

Taken together, our findings show that online Deepfaked content has a strong psychological impact on the viewer, and allows its creator to control public perceptions of others. But how effective they are in doing so? Most, including our own, contain video or audio artefacts, which represent tell-tale signs of manipulation. It's possible that these artefacts undermine the effectiveness of Deepfakes relative to genuine content. Yet, in our studies, this was not the case: Deepfakes were statistically non-inferior to genuine content (i.e., 91% as effective in altering self-reported attitudes (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 intentions compared to genuine content (95% CI [81.4, 117.7]).

It is also worth asking (a) if people are aware that Deepfaking is possible, and (b) if they can detect when they are being exposed to it. Our findings were not encouraging: a large number of participants were unaware that content could be Deepfaked (44%), and even after they were told what it entailed, many were unable to determine if what they had just encountered was genuine or fake. 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, those 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 not the case

in our studies. Aware individuals were manipulated by Deepfakes just as their unaware counterparts were (self-reported attitudes: $\delta = 2.10$, 95% CI [1.83, 2.41], $p < .0001$; implicit attitudes: $\delta = 1.29$, 95% CI [1.03, 1.59], $p < .0001$; intentions: $\delta = 1.51$, 95% CI [1.21, 1.80], $p < .0001$). Those who correctly recognized that they had been exposed to a Deepfake also fell prey to its influence (self-reported attitudes: $\delta = 2.18$, 95% CI [1.93, 2.44], $p < .0001$; implicit attitudes: $\delta = 1.37$, 95% CI [1.12, 1.64], $p < .0001$; intentions: $\delta = 1.59$, 95% CI [1.34, 1.84], $p < .0001$). Deepfakes even biased the attitudes and intentions of those who were both aware that content could be Deepfaked and who had detected that they had been exposed to it (self-reported attitudes: $\delta = 1.98$, 95% CI [1.65, 2.27], $p < .0001$; implicit attitudes: $\delta = 1.35$, 95% CI [1.01, 1.65], $p < .0001$) and intentions ($\delta = 1.38$, 95% CI [1.09, 1.72], $p < .0001$).

In short, even detectable or imperfect Deepfakes can be used to manipulate a viewer's attitudes and intentions, and do so in ways that are similar to authentic content. Many people 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 people from their influence.

Although politicians, journalists, academics, and think-tanks have all warned of the dangers that Deepfakes pose, this research is one of the first to offer systematic empirical support for such concerns. Our results show that a single brief exposure to a Deepfake quickly and effectively shifted (implicit) attitudes and intentions, even when people were fully aware that content can be Deepfaked, and had detected that they had just been exposed to it.

Such findings suggest that technological solutions designed to detect and flag Deepfaked content for viewers will not be enough. What is also needed is a better understanding of the Psychology of Deepfakes, and in particular, how this new technology exploits 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. To examine if these lies root themselves quickly and deeply in our minds, and linger long after efforts to debunk them have ended (21). If so, then corrective approaches currently favored by tech companies, such as tagging Deepfaked content with a warning, may be less effective than currently assumed (22). We also need to examine if Deepfakes can be used to manipulate what we remember, either by installing false memories of events that never happened (known as Mandela effects) or by altering genuine memories that did (23). 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, Deepfakes may cause us to question everything that we see and hear, thereby accelerating a growing trend towards epistemic breakdown: an inability or reduced motivation to distinguish fact from fiction. This "reality apathy" (24) may be exploited by certain actors to dismiss inconvenient or incriminating content (the so-called "liar's dividend" [25]). Given that the human mind is built for belief (26), we may 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 (27). Without such safeguards we may be speeding towards a world where our ability to agree on what is true eventually disappears.

Supplementary materials

Full details of the methods and results can be found in the supplementary materials, along with all preregistrations, data, and code: osf.io/f6ajb

Notes

Author affiliations: S.H. & I.H.: Department of Experimental Clinical and Health Psychology, Ghent University, Belgium; O.F.: Interdisciplinary Center, Herzliya, Israel; M.F.: Department of Psychology, Yale University, USA; C.H.: Fermi National Accelerator Laboratory (Fermilab), USA; R.H.: Rudolf Peierls Centre for Theoretical Physics, Oxford University, UK; X.Y.: Department of Computer Science, Stanford University, USA.

Author contributions: 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. M. Ferguson contributed to study conceptualization, reviewing and editing of 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 designed and implemented the data processing and analyses, contributed to study conceptualization, and reviewed the manuscript.

Competing interests statement: All authors declare we have no competing interests.

Funding: 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 program under the Marie Sklodowska-Curie grant agreement No 722497 - LubISS.

References

1. S. Lewandowsky, U. K. Ecker, J. Cook. Beyond misinformation: Understanding and coping with the "post-truth" era. JARMAC, 6, 353-369. (2017).
2. J. Kietzmann, L. Lee, I. McCarthy, T. Kietzmann, Deepfakes: Trick or treat? Bus. Horiz. 63, 135-146 (2020).
3. R. Satter, Deepfake used to attack activist couple shows new disinformation frontier (2020), (https://www.reuters.com/article/us-cyber-deepfake-activist-idUSKCN24G15E).
4. 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).
5. 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).
6. H. Ajder, G. Patrini, F. Cavalli, L. Cullen, The state of Deepfakes 2019: Landscape, threats, and impact (2019), (https://sensity.ai/reports/).
7. 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/).
8. 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/).
9. T. Hwang, Deepfakes: A grounded threat assessment (Center for Security and Emerging Technology) (2020), (cset.georgetown.edu/research/deepfakes-a-grounded-threat-assessment/).
10. K. Sayler, L. Harris, Deepfakes and national security (2020), (https://crsreports.congress.gov/product/pdf/IF/IF11333).
11. 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).
12. 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).
13. Identifying Outputs of Generative Adversarial Networks Act, S. 2904, 116th Cong., (2019). (https://www.congress.gov/bill/116th-congress/senate-bill/2904).
14. M. Brady, M. Meyer-Resende, Deepfakes: A new disinformation threat (2020), (https://democracy-reporting.org/dri_publications/deepfakes-a-new-disinformation-threat/).
15. 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/).
16. 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/).
17. K. Karsay, D. Schmuck. "Weak, Sad, and Lazy Fatties": Adolescents' Explicit and Implicit Weight Bias Following Exposure to Weight Loss Reality TV Shows. Media Psychol., 22, 60-81. (2019).
18. 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).
19. X. Yao, O. Fried, K. Fatahalian, M. Agrawala, Iterative text-based editing of talking-heads using neural retargeting. arXiv:2011.10688 (2020).
20. A. Mason, How imputations work: The research behind Overdub (2019), (https://blog.descript.com/how-imputations-work-the-research-behind-overdub/).
21. 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).
22. K. Paul, Twitter to label Deepfakes and other deceptive media (2020), (https://www.reuters.com/article/us-twitter-security-idUSKBN1ZY2OV).
23. N. Liv, D. Greenbaum, Deepfakes and memory malleability: False memories in the service of fake news. AJOB Neurosci., 11, 96-104 (2020).
24. A. Ovadya, Deepfake myths: Common misconceptions about synthetic media (2019), (https://securingdemocracy.gmfus.org/deepfake-

myths-common-misconceptions-about-synthetic-media/).
25. B. Chesney, D. Citron, Deepfakes: a looming challenge for privacy, democracy, and national security. Calif. L. Rev., 107, 1753-1819 (2019).
26. N. Porot, E. Mandelbaum, The science of belief: A progress report. WIREs Cog. Sci., 11, 1-17, (2020).
27. S. Van der Linden, J. Roozenbeek, "Psychological inoculation against fake news" in The Psychology of Fake News: Accepting, Sharing, and Correcting Misinformation. R. Greifenader, M. Jaffé, E. Newman, N. Schwarz, Eds. (Psych. Press, 2020).