# **Title**: **Using Deepfakes to Hack the Unconscious Mind**

**Authors**: Sean Hughes1\*, Ohad Fried2, Melissa Ferguson3, David Yao4, Ciaran Hughes5, Rian Hughes6, & Ian Hussey1

**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 Department of Computer Science, Stanford University, USA.

5 Fermi National Accelerator Laboratory (Fermilab), USA.

6 Rudolf Peierls Centre for Theoretical Physics, Oxford University, UK.

\*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 (unconscious) attitudes and intentions. Results show that even imperfect Deepfakes can manipulate viewers, and bias people 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 (unconscious) 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 (*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*). Content generated or manipulated in this way is collectively known as ‘synthetic media’.

Synthetic media is rapidly evolving: it is becoming highly realistic, easier to produce, and thanks to the Internet, can be distributed and shared on a mass scale. One recent report suggests that the number of ‘Deepfakes’ (a subcategory of synthetic media) is doubling online every six months (*5*). What once took a small fortune and a Hollywood special effects department can now be achieved using only a computer or smartphone.

The technology behind synthetic media can be deployed for good or ill. Some are using it to generate believable voices and images for those who have lost their own through traumatic injury and disability (*6*), or to allow celebrities such as David Beckham to deliver public health messages about malaria in nine different languages (*7*). Museums are using it to bring the dead back to life (at the Salvador Dalí Museum visitors can interact with a synthetic Dalí to learn about his art (*8*)), while combining it with natural language processing could one day lead to smart digital assistants capable of truly natural interactions (*9-10*).

However, the technology is also ripe for abuse. Deepfaking has quickly become a tool of harassment against activists (*11*), 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 (*12-13*). Male celebrities are having their voices synthetically copied and digitally distributed (*14*), while their female counterparts are non-consensually grafted into highly realistic pornographic scenes (*5*). Politicians are also being digitally manipulated into endorsing controversial positions (*15*), while worry grows that a well-executed video could have them ‘confess’ to bribery or sexual assault, political disinformation that distorts democratic discourse and election outcomes (*16*).

Unfortunately, the dark side of synthetic media goes even further. Deepfakes have sparked a new disinformation frontier where malicious actors are using the technology to pose as journalists, analysts, or consultants (*17*). These fake identities are legitimized through connections to genuine professionals on LinkedIn (*18*), and used to manipulate mainstream news outlets into publishing content for political or personal gain. 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 (*19-21*). 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.

One day soon it will be impossible to tell with the naked eye or ear if content is genuine or synthetic. Recognizing this inflection point, industry leaders and lawmakers are looking to two forms of protection. Politicians, in Europe and the USA, are advocating for legislation that regulates a technology they believe will further erode the public’s trust in media and push ideologically opposed groups deeper into their own subjective realities (*22-24*). At the same time, technology giants such as Facebook, Google, and Microsoft are developing algorithms to detect Deepfakes, excise them from their platforms, and prevent their spread (*25-26*). Although legislative and technological stopgaps are undoubtedly necessary, they are also in a perpetual game of ‘cat-and-mouse’, with certain actors evolving new ways of evading detection and others rapidly working to catch up. In such a world, no law or algorithm can guarantee that the public will be completely protected from malicious synthetic content.

What is needed then, alongside legislation and technological fixes, is a greater focus on the *human* dimension. It is imperative that we study the impact of this new technology on our thoughts, feelings, and actions. For instance, can Deepfakes be used to manipulate our (unconscious) attitudes and intentions? How effective are they in doing so, especially when compared to authentic 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. We first 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, *δ =* 2.60, 95% CI [2.36, 2.81], *p* < .0000001, unconscious attitudes, *δ =* 1.37, 95% CI [1.17, 1.62], *p* < .0001, and people’s intentions towards the target, *δ =* 2.59, 95% CI [2.37, 2.82], *p* < .0001 (see Fig 1). [[1]](#footnote-2)

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) (*27*), 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 (Experiment 3: self-reported attitudes: *δ =* 2.71, 95% CI [2.57, 2.85], *p* < .0001; unconscious attitudes: *δ =* 1.33, 95% CI [1.19, 1.46], *p* < .0001) (see Fig 2).

Similar findings emerged when a different Deepfake creation method was used, one that generated content from scratch, rather than extracting it from one video and inserting it into another (*28*). This involved taking pre-existing footage from a different actor and using it to generate a 3D head model. 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). 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; unconscious attitudes: *δ =* 1.36, 95% CI [1.14, 1.57], *p* < .0001; behavioral intentions: *δ =* 2.68, 95% CI [2.47, 2.90], *p* < .0001).

The above findings also generalized from one synthetic media type (video) to another (audio). Specifically, 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 (*29*)). 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 a viewer’s attitudes and intentions in ways that were similar to Deepfaked videos, (Experiments 4 and 6: self-reported attitudes: *δ =* 2.71, 95% CI [2.53, 2.88], *p* < .0001; unconscious attitudes: *δ =* 1.32, 95% CI [1.16, 1.49], *p* < .0001; behavioral intentions: *δ =* 3.08, 95% CI [2.69, 3.46], *p* < .0001) (see Fig 2).

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 unconscious attitudes (95% CI [76.1, 121.1]), and 102.6% as effective in altering people’s intentions compared to genuine content (95% CI [92.3, 116.9]).

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 synthetic in nature. That is, they did not make accurate (Balanced Accuracy = .64, 95% CI [.60, 0.67]) nor informed (Youden’s *J* = .27, 95% CI [.20, .35]) 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 (self-reported attitudes: *δ =* 2.10, 95% CI [1.83, 2.41], *p* < .0001; unconscious attitudes: *δ =* 1.29, 95% CI [1.03, 1.59], *p* < .0001; behavioral intentions: 1.29, 95% CI [1.03, 1.59], *p* < .0001) just as their unaware counterparts were. 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; unconscious attitudes: *δ =* 1.37, 95% CI [1.12, 1.64], *p* < .0001; behavioral intentions: 2.57, 95% CI [2.33, 2.83], *p* < .0001). Deepfake even changed the attitudes (self-reported: *δ =* 1.98, 95% CI [1.65, 2.27], *p* < .0001; unconscious attitudes: *δ =* 1.35, 95% CI [1.01, 1.65], *p* < .0001) and intentions (*δ =* 2.39, 95% CI [2.07, 2.70], *p* < .0001), of those who were aware *and* detected its presence.

In short, even detectable or imperfect Deepfakes psychologically influence many viewers, and can be used to manipulate attitudes and intentions just as effectively as authentic 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 served to protect individuals from its influence.

Given the dangers posed by malicious synthetic content, politicians are looking to the law to help regulate its creation and spread, while industry leaders invest in technological solutions to help consumers detect and recognize when they are exposed to it. Our findings suggest that this may not be enough: a single brief exposure to a Deepfake quickly and effectively shifted (unconscious) 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. We need to examine if these lies root themselves quickly and deeply in our minds, and linger on long after efforts to debunk them have ended (as is the case with fake news (*30-31*)).

If so, then corrective approaches currently favored by tech companies, such as tagging Deepfaked videos with a warning, may be less effective than assumed (*32*). 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 (*33*). 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” (*34*) may be exploited by certain actors to dismiss inconvenient or incriminating content as a fabrication (the so-called ‘liar’s dividend’ (*35*)). Given that the human mind is built for belief (*36*), we need psychological interventions that can inoculate individuals against synthetic media attacks, 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 individual and collective 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), (available at <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), (available at <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), (available at <https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3>).

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

6. J. Cattiau, How Tim Shaw regained his voice (2019), (available at <https://www.blog.google/outreach-initiatives/accessibility/how-tim-shaw-regained-his-voice/>).

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

8. D. Lee, Deepfake Salvador Dalí takes selfies with museum visitors (2019), (available at <https://www.theverge.com/2019/5/10/18540953/salvador-dali-lives-deepfake-museum>).

9. T. Young, D. Hazarika, S. Poria, E. Cambria, Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, ***13***, 55-75 (2018).

10. B. Chen, C. Metz, Google’s Duplex uses A.I. to mimic humans (sometimes) (2019), (available at <https://www.nytimes.com/2019/05/22/technology/personaltech/ai-google-duplex.html>).

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

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

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

14. B. Hochberg, YouTube won’t take down a Deepfake of Jay-Z reading Hamlet - To sue, or not to sue (2020), (available at <https://www.forbes.com/sites/williamhochberg/2020/05/18/to-sue-or-not-to-sue---that-is-the-jay-zs-deepfake-question>).

15. J. Koetsier, Fake video election? Deepfake videos ‘grew 20X’ since 2019 (2020), (available at <https://www.forbes.com/sites/johnkoetsier/2020/09/09/fake-video-election-deepfake-videos-grew-20x-since-2019/>).

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

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

18. R. Satter, Experts: Spy used AI-generated face to connect with targets (2019), (available at <https://apnews.com/article/bc2f19097a4c4fffaa00de6770b8a60d>).

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

20. K. Sayler, L. Harris, Deepfakes and national security (2020), (available at <https://crsreports.congress.gov/product/pdf/IF/IF11333>).

21. 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), (available at <https://www.europol.europa.eu/publications-documents/malicious-uses-and-abuses-of-artificial-intelligence>).

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

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

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

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

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

27. 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).

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

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

30. 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).

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

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

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

34. A. Ovadya, Deepfake myths: Common misconceptions about synthetic media (2019), (available at <https://securingdemocracy.gmfus.org/deepfake-myths-common-misconceptions-about-synthetic-media/>).

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

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

**Acknowledgments**

This research was supported by XXX. 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 D. 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.

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. [↑](#footnote-ref-2)