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 (automatic) attitudes as well as behavioral intentions. Results indicated that Deepfaked videos and audio can be used to control public perceptions of others, and are just as effective in doing so 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 them from its influence. All preregistrations, data, and code are 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 manipulated. A politician or celebrity's comments can be edited and misreported while images on magazine covers, advertisements, and websites can be altered to depict their contents as being better than they actually are. In contrast, we are relatively less inclined to think of video and audio recordings as easy to manipulate, and instead assume that they provide accurate and valid information about others. Put simply, seeing is still very much believing.

However, this may no longer be true. A branch of artificial intelligence 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) (see https://www.youtube.com/watch?v=cQ54GDm1eL0).

Deepfakes are rapidly evolving: they are 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' is doubling online every six months (3). What once took a small fortune and a Hollywood special effects department can now be achieved using only a computer or smartphone.

Deepfaking has quickly become a tool of harassment against activists (4), 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 (5-6). Female celebrities are being Deepfaked into highly realistic pornographic scenes (7), 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 (8-9). 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 (10-12).

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 (13-15). 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 (16-17).

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. With the above in mind, we set out to examine the following questions. Is a single brief exposure to a Deepfake enough to manipulate our (automatic) attitudes and behavioral intentions towards others? Just how effective are they in influencing viewers, especially when compared to authentic online content? Are people aware that Deepfaking is even possible, and perhaps more importantly, can they detect when they are being exposed to one? Finally, does an awareness of Deepfaking and the ability to detect when it is present immunize them from its influence?

Results

Experiments 1a-1b: Genuine online content can be used to engineer public perceptions of a target.

We carried out seven preregistered studies (N = 2558) to answer these questions. In Experiments 1a and 1b we wanted to know if public perceptions (i.e., self-reported and automatic attitudes) can be manipulated by selectively exposing people to video recordings of a target individual. We therefore started with authentic videos given that these are easiest to create and most ubiquitous on the internet. In these videos we had a novel individual ('Chris') disclose personal information about himself. In one video, he uttered three highly positive self-statements while in another video he uttered three highly negative statements (both videos also included two neutral statements). Participants navigated to YouTube and either watched the positive or negative video variant. Thereafter their self-reported and automatic attitudes were assessed.

Consistent with past work (18), our first two studies show that genuine online content leads to social learning at both the implicit and explicit levels. Self-reported attitudes strongly differed depending on the type of content people encountered (Experiment 1a: $t(145.74) = 14.98$, $p < .001$, $d = 2.46$, 95% CI [2.03; 2.89], $BF_{10} > 10^5$; Experiment 1b: $t(129.94) = 15.73$, $p < .001$, $d = 2.71$, 95% CI [2.24; 3.18], $BF_{10} > 10^5$). Manipulating the information environment allowed us to control public sentiment towards the target, having him be liked by some and despised by others. A similar finding emerged at the implicit level, with automatic attitudes also strongly differing depending on the type of content encountered (Experiment 1a: $t(138.23) = 8.23$, $p < .001$, $d = 1.35$, 95% CI [0.99; 1.71], $BF_{10} > 10^5$; Experiment 1b: $t(126.9) = 7.78$, $p < .001$, $d = 1.35$, 95% CI [0.97; 1.73], $BF_{10} > 10^5$). Our findings confirm that implicit and explicit attitudes towards a novel target can be engineered via selective exposure to authentic online content.

Experiment 2: Deepfaked videos are highly effective in manipulating public perceptions of others (*'Cut and paste' method*).

In Experiment 2 we set out to replicate our findings with one important addition. This time half of participants watched an authentic video of the target whereas the other half watched a Deepfake of that same individual. At its core, this Deepfake constituted a digital copy of the target that provided us with total control over his appearance and actions. This allowed us to manipulate public perceptions towards him with the same level of precision as in Experiments 1a-1b.

Deepfakes were created using the 'cut-and-paste' method (*see below*). Broadly speaking, this involves 'cutting' a target's genuine content from context A and 'pasting' it into an entirely different video of them in context B. The use cases of this technique are manifold. For instance, it can be used to extract statements made by a political candidate from one video (e.g., where they talk about the dangers that climate change pose), modify them, and insert them into a completely different video, where they now appear to talk about the dangers that racial outgroups pose. This same method could also be used to 'scrape' publically available content from a person's social media account and Deepfake them for malicious (blackmailing) purposes.

In our case we took the genuine videos from Experiment 1b, fitted a parameterized 3D model to the target's head, and then used this model to generate computer graphical renderings of his face and mouth movements. These renderings were then converted to photorealistic synthesized video using a trained Generative Adversarial Network (19) 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. This allowed us to investigate if (a) Deepfaked online content could be used to manipulate

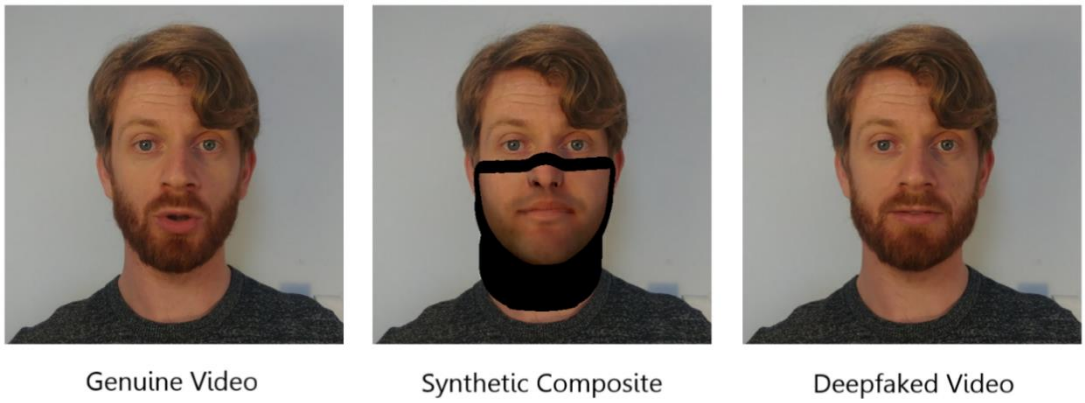

Figure 1. Deepfake creation method (‘*Fabricate from scratch*’). This approach leverages a small amount of the target’s genuine data as well as a large repository of speaking footage of a different individual to generate high quality 3D head model parameters for the desired Deepfaked content. This approach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

perceptions of the target and if (b) such perceptions were as strong as those established via genuine content.

Results showed that by selectively exposing people to positive or negative information we could control how the target was publically perceived, replicating our earlier findings. This was true for both self-reported attitudes, $t(318.43) = 20.62$, $p < .001$, $d = 2.22$, 95% CI [1.96; 2.49], $BF_{10} > 10^5$, and implicit attitudes, $t(317.27) = 9.92$, $p < .001$, $d = 1.07$, 95% CI [0.85; 1.29], $BF_{10} > 10^5$. More importantly, we found that Deepfakes successfully manipulated public perceptions of the target, and did so in ways that were similar to authentic content, both at the explicit, $t(355.83) = -0.10$, $p = .92$, $d = 0.01$, 95% CI [-0.22; 0.20], $BF_{10} = 0.12$, and implicit levels, $t(353) = 0.52$, $p = .60$, $d = 0.06$, 95% CI [-0.15; 0.26], $BF_{10} = 0.13$ (see Fig. 2).

Experiment 3: Deepfaked videos are highly effective in manipulating public perceptions of others (*‘Fabricate from scratch’ method*).

Deepfakes can be created in many ways. For instance, in Experiment 2, we took pre-existing content from context A and digitally grafted it into context B, thereby having the target confess to actions he had never previously committed. But what about a situation where the desired content doesn’t actually exist and has to instead be fabricated from scratch. Could this alternative creation process also be used to control public perceptions of the target? Experiment 3 examined this idea. Participants were asked to complete a similar procedure as in Experiment 2. This time, however, we took advantage of a newly developed method by Yao et al. (20) to generate the Deepfakes. 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 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 videos were similar in their content but differed in their origin (i.e., genuine vs Deepfaked; see Fig. 1).

Digitally manipulating the targets actions in this way allowed us to once more influence the viewer’s thoughts and feelings (self-reported attitudes: $t(212.9) = 17.12$, $p < .001$, $d = 2.31$, 95% CI [1.97; 2.66], $BF_{10} > 10^5$; implicit attitudes: $t(212.04) = 9.34$, $p < .001$, $d = 1.26$, 95% CI [0.97; 1.55], $BF_{10} > 10^5$). Once again, we found that Deepfaked content successfully manipulated public perceptions, and did so in ways that were similar to authentic content, both at the explicit, $t(218.79) = -1.01$, $p = .32$, $d = -0.14$, 95% CI [-0.39; 0.13], $BF_{10} = 0.24$, and implicit levels, $t(216.69) = 0.95$, $p = .35$, $d = 0.13$, 95% CI [-0.14; 0.39], $BF_{10} = 0.22$. Experiment 3 therefore replicated our core findings and generalized them from one Deepfake creation process (i.e., where pre-existing content is digitally grafted from context A into context B) to another (i.e., where the desired content was created from scratch and used to fabricate a malicious or flattering video of the target).

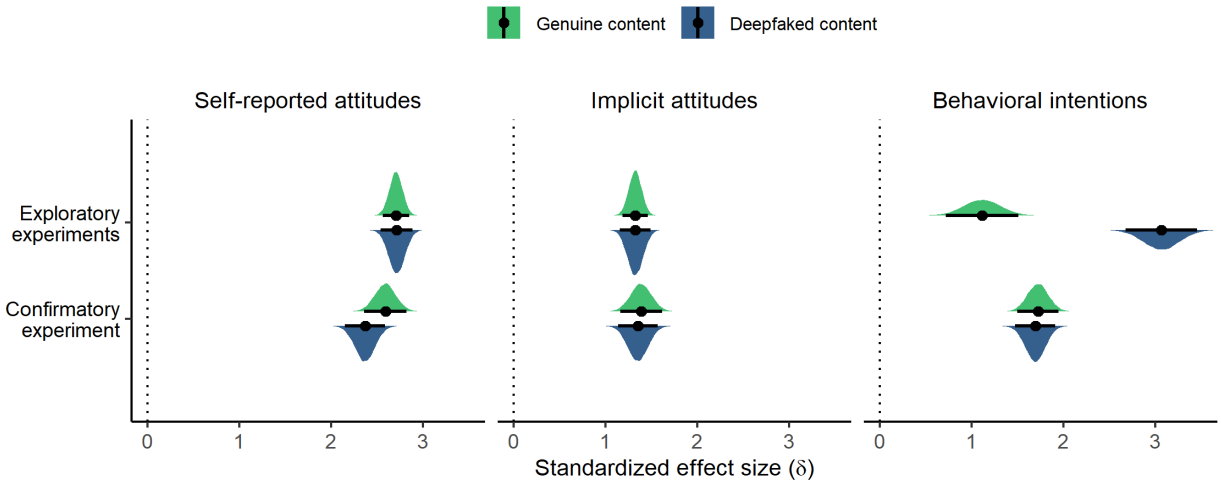

Figure 2. Standardized effect sizes, 95% confidence intervals, and distributions for self-reported attitudes, implicit attitudes, and behavioral intentions for those exposed to genuine and Deepfaked online content. 'Exploratory experiments' refers to combined effects from Experiments 1-5 while 'Confirmatory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

Experiments 4 and 5: Deepfaked audio is also highly effective in manipulating public perceptions of others.

So far we have relied on Deepfaked *videos* to influence public perceptions of the target. But could a similar outcome be achieved through audio alone? If so, then this would provide a far cheaper, less resource intensive, and widely available method of attitude and behavior manipulation than Deepfaked videos. Indeed, it appears this technique has already seen real-world use: in one recent case hackers Deepfaked a CEO's voice and used it to trick an employee into initiating a six-figure wire transfer (21).

Experiments 4-5 set out to examine if our findings would generalize from one Deepfake type (video) to another (audio). A similar procedure was used as in our previous studies but with one notable change: videos were now replaced with audio clips. Genuine audio clips were created by extracting audio from the videos used in Experiments 2. Deepfaked audio clips were generated by first creating a training set of Chris's voice and then feeding that information to a bidirectional text-to-speech autoregressive neural network. This process enabled the neural network to learn how to closely mimic Chris's voice (22). The end result was an entirely Deepfaked voice: a synthetic replica that sounded similar to the original, and which could be manipulated into saying anything. In our case we used it to create the statements, and thus the positive and negative audio clips, for these two studies.

Participants were informed that we were interested in how they remember and react to online information. Their task was to listen to an audio recording of a person called Chris that was extracted from his YouTube channel and to answer questions about what they had just heard. Thereafter they listened to either a positive or negative audio variant that was authentic or Deepfaked in nature, and then completed the various attitude measures.

Results show that synthetically cloning the target's voice and manipulating what he 'said' gave us control over how he was perceived by others, both at the explicit level (self-reported attitudes; Experiment 4: $t(330.86) = 25.92$, $p < .001$, $d = 2.81$, 95% CI [2.51; 3.11], $BF_{10} > 10^5$; Experiment 5: $t(186.84) = 20.91$, $p < .001$, $d = 2.89$, 95% CI [2.51; 3.28], $BF_{10} > 10^5$) and implicit levels (automatic attitudes; Experiment 4: $t(335.69) = 11.18$, $p < .001$, $d = 1.21$, 95% CI [0.98; 1.44], $BF_{10} > 10^5$; Experiment 5: $t(200.89) = 9.93$, $p < .001$, $d = 1.36$, 95% CI [1.06; 1.66], $BF_{10} > 10^5$). Deepfaked audio successfully biased public perceptions of the target, and gave rise to self-reported attitudes of similar magnitude as authentic audio in Experiment 4, $t(335.41) = 1.09$, $p = .28$, $d = 0.12$, 95% CI [-0.10; 0.33], $BF_{10} = 0.21$, and even larger attitudes than authentic audio in Experiment 5, $t(206.7) = 2.92$, $p = .004$, $d = 0.39$, 95% CI [0.13; 0.67], $BF_{10} = 7.95$. Deepfaked audio also installed implicit attitudes of similar magnitude as genuine audio (Experiment 4: $t(337.26) = -0.37$, $p = .71$, $d = -0.04$, 95% CI [-0.25; 0.17], $BF_{10} = 0.13$; Experiment 5: $t(216) = -0.18$, $p = .85$, $d = -0.03$, 95% CI [-0.29; 0.24], $BF_{10} = 0.15$).

Experiment 6. Pre-registered, high-powered, confirmation study.

Experiments 1-5 demonstrate that Deepfakes can quickly and powerfully impact viewers, equipping their creators with a ready means of controlling public perceptions of others. This is true for different types of Deepfaked content (video and audio) and different Deepfake creation methods ('cut and paste' vs. 'fabricate from scratch').

Experiment 6 represents a high-powered, preregistered, confirmation of our findings. In several cases, our prior hypotheses had already been strongly supported by evidence from preregistered analyses in Experiments 1-5 (e.g., can genuine and Deepfaked online content be used to engineer public perceptions of others). In other cases, we had new hypotheses that were either induced from, or refined based on, our initial data and thus required confirmation. For instance, we were curious to know if Deepfakes could also be used to alter behavioral intentions towards a target. Similarly, are people aware that Deepfaking is even possible, and are they able to detect when they are being exposed to such forgeries? Does a general awareness that content can be Deepfaked increase one's chances of detecting it, and does awareness or detection help 'immunize' them from its influence?

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; for more on these changes see Supplementary Materials and Appendix A: Meta-Analysis). Participants completed a similar procedure as in Experiment 3 that also included behavioral intentions, Deepfake awareness, and Deepfake detection measures. We consider each of our research questions separately below.

Research Question 1: Can Deepfakes be used to manipulate public perceptions of others? Results confirmed that manipulating the informational content of genuine videos (i.e., positive vs. negative statements) influenced self-reported attitudes (Standardized effect size $\delta = 2.60$, 95% CI [2.36, 2.81], $p < .0000001$), and implicit attitudes ($\delta = 1.37$, 95% CI [1.17, 1.62], $p < .0000001$), as well as behavioral intentions ($\delta = 1.74$, 95% CI [1.50, 1.95], $p < .0000001$). The same was true for Deepfaked content, which also influenced self-reported attitudes ($\delta = 2.35$, 95% CI [2.15, 2.59], $p < .0000001$), implicit attitudes ($\delta = 1.36$, 95% CI [1.14, 1.57], $p < .0000001$), and behavioral intentions ($\delta = 1.70$, 95% CI [1.48, 1.91], $p < .0000001$).

Research Question 2: Are Deepfakes as effective in engineering public perceptions as genuine content? It's not only important to know that Deepfakes can manipulate attitudes and intentions. We also need to know 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]).

Research Question 3: Are people aware that content can be Deepfaked and how effective are they in detecting when they are being exposed to it? It is also worth asking if (a) people are aware that Deepfaking is possible, and if (b) 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]).

Research Question 4: Does prior awareness of Deepfaking (or an ability to detect when it is present) help immunize people from its influence? 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], $p < .0001$).

Research Question 5: Does awareness & detection better protect one from the influence of Deepfakes? People who are aware of Deepfaking and who say that they detected its presence should (arguably) be the most likely to resist its influence. Unfortunately, we found that 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], $p < .0001$).

Taken together, our findings confirm that 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.

Discussion

Although politicians, journalists, academics, and think-tanks have all warned of the dangers that Deepfakes pose, our paper is one of the first to offer systematic empirical support for those concerns. Our results show that a single brief exposure to a Deepfake quickly and effectively shifted (implicit) attitudes and intentions towards a target, 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 (23). 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 (24). 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 (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, 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" (26) may be exploited by certain actors to dismiss inconvenient or incriminating content (the so-called "liar's dividend" [27]). Given that the human mind is built for belief (28), 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 (29). Without such safeguards we may be speeding towards a world where our ability to agree on what is true eventually disappears.

Method

Participants and Design

165 participants (92 male, *M*age = 30.4, *SD* = 7.6) [Experiment 1a], 167 participants (91 female, *M*age = 31.5, *SD* = 7.6) [Experiment 1b], 428 participants (232 female, *M*age = 30.7, *SD* = 9.0) [Experiment 2], 276 participants (151 female, *M*age = 32.6, *SD* = 12.3) [Experiment 3], 429 participants (258 female, *M*age = 30, *SD* = 8.6) [Experiment 4], 265 participants (154 female, *M*age = 33.3, *SD* = 12.6) [Experiment 5], and 828 participants (476 female, *M*age = 35.9, *SD* = 13) [Experiment 6] took part via the Prolific website (https://prolific.ac) in exchange for a monetary reward. Assignment to the different *Information Content* conditions (positive or negative behavioral statements) was counterbalanced across participants in all studies. Assignment to the *Information Type* conditions (Genuine vs. Deepfaked) was counterbalanced across participants in Experiments 2-6.

Ratings and IAT scores were the dependent variables. One method factor was also counterbalanced across participants: *evaluative task order* (whether participants encountered the self-report ratings or IAT first). Study designs and data-analysis plans for all experiments are available on the Open Science Framework website (https://osf.io/f6ajb/). We report all manipulations and measures used in our experiments. All data were collected without intermittent data analysis. The data analytic plan, stimuli and materials, experimental scripts, and data are available at the above link. Deviations from pre-registration can also be found at the above link.

Stimuli

Attitude Objects. A novel individual (Chris) served as the target during the attitude induction phase (this individual was the first author who was selected on the basis of convenience). Chris appeared during the video or audio while his images also served as one set of category stimuli during the IAT. A second individual (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.

Behavioral Statements. Eight behavioral statements were selected for use in the videos and audio: three positive, three negative, and two neutral. These items were selected from a larger pool 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). See the Supplementary Materials for the statements used in Experiments 1-6.

Personalized IAT (pIAT). A set of eight positive and eight negative trait adjectives were used as valenced stimuli during the pIAT in Experiments 1-5. 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. Note: only the first five positive and negative adjectives were used in Experiment 6.

Procedure

Participants were welcomed to the study and asked for their informed consent. Studies generally consisted of four sections: demographic information, attitude formation phase, attitude assessment measures, and exploratory questions. Thereafter participants were thanked and debriefed.

Demographics

Participants were asked to self-report their age and gender in Experiments 1-6, and to report their country of residence, ethnicity, educational level, employment status, and income in Experiments 3 and 5.

Attitude Formation Phase

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 [listen to audio: Experiments 4-5] taken from a YouTube channel. The person who makes these videos [audio] is called Chris. Please watch Chris' video [listen to Chris' audio] and pay close attention to what he says. We will ask you questions about this later on."

Thereafter the experiment played an embedded YouTube video [audio] of Chris. During the video [audio], Chris emitted three valenced statements as well as two neutral statements about himself. Half of the participants encountered positive variant video/audio wherein Chris emitted three positive and two neutral statements, whereas the other half encountered a negative variant video/audio, wherein Chris emitted three negative and two neutral statements. In Experiments 1a-1b the content was authentic, whereas in Experiments 2-6 the content was either authentic or Deepfaked in nature (see https://osf.io/f6ajb/ for the videos and audio used in Experiments 1-6).

Attitude Assessment Measures

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) 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: 20 and 40 trials were used in place of the 16 and 32 trial structure in Experiment 6.

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.

Behavioral Intentions. In Experiments 5-6 participants were asked to indicate how they intended to behave with respect to the target ("1. If I were browsing YouTube and encountered Chris' video, I would support him by clicking the 'share' button (i.e., share his video with other people)"; "2. Chris has just started to make these videos and wants to become a YouTuber. I happen to encounter his video on YouTube. I would 'subscribe' to his channel to learn more about him." "3. I would recommend Chris' videos to others"). In Experiment 5 responds were emitted using a scale ranging from -2 (*Strongly Disagree*) to 2 (*Strongly Agree*) with 0 (Neutral) as a center point. In Experiment 6 the scale ranged from -3 to +3.

Deepfake Detection. Participants in Experiment 6 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 6.

Deepfake Awareness. Prior awareness of Deepfaking as a concept was assessed in Experiment 6 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 6. Note: Deepfake awareness and detection were also probed in Experiment 3.

Individual Difference Measures

A number of individual difference measures were taken in Experiment 3, including measures of political ideology, religiosity, cognitive ability (revised cognitive reflection test [rCRT]), preference for effortful or intuitive thinking styles (rational-experiential inventory [REI]), overclaiming, conspiratorial thinking, and deepfake awareness and detection. Preference for effortful vs. intuitive thinking (REI), and cognitive ability (rCRT) were also taken in Experiment 5. However, 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) (See Supplementary Materials for additional information on each of these measures).

Note: 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 were not analyzed in this paper (but simply reported for transparency purposes). We have made all data and analyses related to demographic and individual difference factors available to others who are interested in such questions (see https://osf.io/f6ajb/).

Exploratory Questions

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

Data Analysis

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 1a], n = 32 [Experiment 1b], n = 70 [Experiment 2], n = 55 [Experiment 3], n = 88 [Experiment 4], n = 47 [Experiment 5]). We also excluded data in Experiment 6 if participants spent too little (minimum of 2.25 minutes) or too much time on the attitude induction phase (over 4.5 minutes watching the video) ($n = 192$).This led to a final sample of 148 participants in Experiment 1, 135 in Experiment 1b, 358 in Experiment 2, 221 in Experiment 3, 341 in Experiment 4, 218 in Experiment 5, and 635 in Experiment 6.

Data Preparation (Exploratory Studies 1-5)

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

Data Preparation (Confirmatory Study 6)

Data were prepared as noted above. However, similar to our meta-analyses (see Appendix A), 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 *Information Content* condition [positive vs. negative], and by *Information Type* [Authentic vs. Deepfaked]).

Analytic Strategy (Exploratory Studies 1-5)

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 *Information Content* (positive vs. negative behavioral statements) (independent variable). A series of independent and one-sample *t*-tests were also carried out on the ratings and pIAT data to determine if they differed as a function of *Information Type* (authentic vs. Deepfaked) in Experiments 2-5. 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 Information Content and/or Information Type (alternative hypothesis) or that there is no such difference (null hypothesis).

Analytic Strategy (Confirmatory Study 6)

A similar analytic strategy was employed as outlined in the Meta-Analysis (see Appendix A). 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).

Supplementary Materials

Details of the materials used in Experiments 1-6 can be found in the Supplementary Materials at osf.io/muvte. A detailed overview of the meta-analysis can be found in Appendix A at https://osf.io/h5v98/. All preregistrations, data, and code can be found at 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. H. Ajder, G. Patrini, F. Cavalli, L. Cullen, The state of Deepfakes 2019: Landscape, threats, and impact (2019), (available at https://sensity.ai/reports/).
4. R. Satter, Deepfake used to attack activist couple shows new disinformation frontier (2020), (https://www.reuters.com/article/us-cyber-deepfake-activist-idUSKCN24G15E).
5. 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).
6. 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).
7. H. Ajder, G. Patrini, F. Cavalli, L. Cullen, The state of Deepfakes 2019: Landscape, threats, and impact (2019), (https://sensity.ai/reports/).
8. 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/).
9. 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/).
10. T. Hwang, Deepfakes: A grounded threat assessment (Center for Security and Emerging Technology) (2020), (cset.georgetown.edu/research/deepfakes-a-grounded-threat-assessment/).
11. K. Sayler, L. Harris, Deepfakes and national security (2020), (https://crsreports.congress.gov/product/pdf/IF/IF11333).
12. 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).

13. 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).
14. Identifying Outputs of Generative Adversarial Networks Act, S. 2904, 116th Cong., (2019). (https://www.congress.gov/bill/116th-congress/senate-bill/2904).
15. M. Brady, M. Meyer-Resende, Deepfakes: A new disinformation threat (2020), (https://democracy-reporting.org/dri_publications/deepfakes-a-new-disinformation-threat/).
16. 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/).
17. 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/).
18. 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).
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. Stupp, C. Fraudsters used AI to mimic CEO's voice in unusual cybercrime case. The Wall Street Journal, 30(08) (2019).
22. A. Mason, How imputations work: The research behind Overdub (2019), (https://blog.descript.com/how-imputations-work-the-research-behind-overdub/).
23. 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).
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).
29. 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).