Deepfaked Online Content is Highly Effective in Manipulating Attitudes & Intentions

Abstract

Disinformation has spread rapidly through social media and news sites, biasing our (moral) judgements of individuals and groups. “Deepfakes”, a new type of AI-generated media, represent a powerful tool for spreading disinformation online. Although they may appear genuine, Deepfakes are hyper-realistic fabrications that enable one to digitally control another person’s appearance and actions. Across seven preregistered studies (N = 2558) we examined the psychological impact of Deepfakes on viewers. Participants were exposed to genuine or Deepfaked online content, after which their (implicit) attitudes and behavioral intentions were measured. We found that Deepfakes are highly effectively in manipulating public perceptions, and do so in ways that are similar to genuine content. Many people are unaware that Deepfaking is possible, find it difficult to detect when they are being exposed to it, and neither awareness nor detection serves to protect them from its influence. Preregistrations, data, and code are available at osf.io/f6ajb.

Keywords: Deepfakes, AI-Generated Media, Implicit, Attitudes, Intentions, Public Perceptions

Statement of Relevance

Conventional wisdom dictates that seeing is believing. However, thanks to recent advances in artificial intelligence, this may no longer be the case. A branch of machine learning known as 'deep learning' has made it increasingly easy to take a person's face, voice, or writing style, feed that data to a computer algorithm, and have it generate a synthetic copy. This 'Deepfake' can be used to convince others that what they are seeing, reading, or hearing is fact rather than fiction. Concern grows that Deepfakes pose a danger for the business, entertainment, intelligence, and political sectors. Yet the psychological impact of this new technology has never been systematically investigated. Across seven studies we exposed people to Deepfaked or genuine online content. Results show that Deepfakes are highly effective in manipulating public perceptions, and give rise to attitudes that are just as strong as those established by authentic content.

Deepfaked Online Content is Highly Effective in Manipulating Attitudes & Intentions

The proliferation of social media, news, and gossip sites, has brought with it an 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 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 (Lewandowsky, Ecker, & Cook, 2017).

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 easily manipulable, 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.

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 (Kietzmann, Lee, McCarthy, & Kietzmann, 2020). Indeed, one report suggests that the number of 'Deepfakes' is doubling online every six months (Ajder, Patrini, Cavalli, & Cullen, 2019). What once took a small fortune and a Hollywood special effects department can now be achieved using only a computer or smartphone.

Deepfakes have quickly become a tool of harassment against activists (Satter, 2020), 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 (Bateman, 2020; Stupp, 2020). Female celebrities are being inserted into highly realistic pornographic scenes (Ajder et al., 2019), while worry grows that a well-executed video could have a politician 'confess' to bribery or sexual assault, disinformation that distorts democratic discourse and election outcomes (Galston, 2020; Koetsier, 2020). Elsewhere, intelligence services and think tanks warn that Deepfakes represent a growing cybersecurity threat, a tool that state-sponsored actors, political groups, and lone individuals could use to trigger social unrest, fuel diplomatic tensions, and undermine public safety (Hwang, 2020; Sayler & Harris, 2020; Ciancaglini et al., 2020).

Recognizing these dangers, politicians in Europe and the USA have called for legislation to regulate a technology they believe will further erode the public's trust in media, and push ideologically opposed groups deeper into their own subjective realities (EU Commission, 2018; Cortez Masto, 2019). At the same time, industry leaders such as Facebook, Google, and Microsoft are developing algorithms to detect Deepfakes, excise them from their platforms, and prevent their spread (Burt & Horvitz, 2020; Canton Ferrer et al., 2020).

Although legislative and technological stopgaps are undoubtedly necessary, they are also in a perpetual game of 'cat-and-mouse', with certain actors evolving new ways of evading detection and others rapidly working to catch up. In such a world, no law or algorithm can guarantee that the public will be completely protected from malicious synthetic content.

What is needed then, alongside legislation and technological fixes, is a greater focus on the *human* dimension. It is imperative that we start studying the impact of this new technology on our thoughts, feelings, and actions. For instance, can a single brief exposure to Deepfaked online content manipulate our (implicit) attitudes and behavioral intentions? Just how effective are Deepfakes in biasing 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?

The Current Research

We carried out seven preregistered studies (N = 2558) to answer these questions. Experiments 1a and 1b examined if *genuine* online content could be used to manipulate public perceptions of others. We initially focused on genuine videos because they are one of the most common media types online and provide a benchmark against which the effectiveness of Deepfakes can be directly compared. In these videos a target individual disclosed information about himself. In one video, he uttered three positive self-statements while in another video he uttered three negative statements (both videos also included two neutral statements). Participants navigated to YouTube, watched one of these videos, and provided measures of their self-reported and automatic attitudes. Briefly, genuine online content led to social learning at both the implicit and explicit levels.

Experiment 2 replicated this design but with one key addition. This time participants either watched a genuine video of the target or a Deepfake of that same individual. Deepfakes were created using the popular 'cut-and-paste' method: extracting ('cutting') a target's words and actions from context A and inserting ('pasting') them into an entirely different video in context B (Fried et al., 2019).[1] We used this technique to have the target 'confess' to either virtuous or malicious actions he had never previously committed. By strategically deploying these Deepfakes we had total control over how he was perceived, liked by some and despised by others. More importantly, Deepfakes led to implicit and explicit attitudes that were just as strong as those established when one views authentic online content.

[1] This technique is ripe for abuse. It can be used to extract statements made by a political candidate from one video (e.g., where they genuinely talk about the dangers that climate change pose), modify them, and insert them into a completely different video, where they now appear to warn about the dangers that racial outgroups pose. It can also be used to 'scrape' publically available content from a person's social media account and Deepfake them for malicious (blackmailing) purposes.

Experiment 3 simulated an even more realistic scenario, one where pre-existing footage of a target does not exist and has to be completely fabricated (Yao et al., 2020). This technique can be used to place words in the mouth of a political opponent, discredit activists, or for purposes of identity theft or impersonation (e.g., Bateman, 2020). This form of Deepfaking was also highly effective in influencing a viewer's thoughts and feelings.

The above studies relied on Deepfaked *videos* to influence public perceptions. But can a similar outcome be achieved using audio alone? If so, then this would provide a cheaper, less resource intensive, and more widely available method of attitude and behavior manipulation than video content. For instance, hackers recently Deepfaked a CEO's voice and used it to trick an employee into initiating a six-figure wire transfer (Stupp, 2019).

Experiments 4-5 examined this possibility. A similar procedure was used as before but with one key difference: videos were replaced with audio clips. Genuine clips were created by extracting audio from the authentic videos used in Experiment 3. Deepfakes were generated by creating a training set of the target's voice and then feeding that information to a bidirectional text-to-speech autoregressive neural network (Mason, 2019). This process allowed the neural network to learn how to mimic his voice. The end result was an entirely Deepfaked voice: a synthetic replica that sounded similar to the target, and which could be manipulated into saying anything. Cloning the target's voice and manipulating what he 'said' gave us similar control over how he was perceived as Deepfaked videos.

Our final study (Experiment 6) served as a high-powered, pre-registered, confirmation of the above findings. In several cases, our prior hypotheses had been strongly supported (e.g., can genuine and Deepfaked online content be used to engineer public perceptions). In other cases, we had new hypotheses that were either induced from, or refined based on, our initial data and therefore required confirmation. Improvements were made to the design, preregistration specificity, and analytic strategy (e.g., we swapped to a Bayesian framework to

produce more intuitive effect sizes and tests of non-inferiority). Participants completed a similar procedure as in Experiment 3 that also included behavioral intentions, Deepfake awareness and detection measures. We found that even detectable or imperfect Deepfakes can be used to manipulate a viewer's attitudes and intentions, and do so in ways that are comparable to genuine content. Many are unaware of this new technology, find it difficult to detect when they are being exposed to it, and neither awareness nor detection served to protect them from its influence. We consider the implications of our findings in the General Discussion.

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.

Method

Sample Size Selection

Samples were selected on a convenience basis for the exploratory studies (Experiment 1-5). For the high-powered, pre-registered confirmation study (Experiment 6) size was determined via Bayesian power analysis which was itself determined using a simulation study. The simulation involved the following steps. Bayesian linear models were first fitted to the data from Experiments 1-5 to provide point estimates of the parameters used in these hypothesis tests. These parameters were then used to simulate data that met the same 'true'

parameters. The models were then refit to the simulated data, and hypothesis tests were applied. 1000 iterations of this "simulate-data-fit model-test hypotheses" process were then performed. The proportion of simulations which detected the known 'true' effects (i.e., statistical power) was then summarized. The number of participants simulated was varied between simulation runs until a sample size was obtained that provided at least 80% power for all hypotheses. This sample size was then adjusted to take the data exclusion rates observed in Experiments 1-5 into account. Results indicated that 600 participants would be required after exclusions.

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 counterbalanced across participants (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, materials, experimental scripts, data, and deviations from pre-registration are available at the above link.

Stimuli

Attitude Objects. A novel individual (Chris) served as the target during the attitude formation phase (this individual was the first author who was selected on the basis of convenience). The target appeared during the video or audio while his images also served as one set of category stimuli during the pIAT. A second individual (Bob) was selected from a large face database and served as the contrast category during the IAT. '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 that were pre-tested along three dimensions: valence, believability, and diagnosticity (see 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. 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.

Authentic Content. Authentic positive videos were created using three positive behavioral statements and two neutral statements, while negative videos were created using three negative and two neutral statements. Authentic audio clips were created by extracting the audio content from the authentic videos used in Experiment 3.

Deepfaked Content. In Experiment 2, Deepfakes were created using the 'cut-and-paste' method. 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. 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 (Fried et al., 2019) and served as the raw input for the Deepfakes. Specifically, a Deepfaked negative video was created by ‘cutting’ the positive statements out of authentic positive videos and ‘pasting’ in the Deepfaked negative statements, while a Deepfaked positive video was created by cutting the negative statements from the authentic negative video and pasting in the Deepfaked positive statements. These fabricated videos were then uploaded to YouTube where participants watched them.

In Experiment 3 we took advantage of a newly developed method by Yao et al. (2020) 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 (see Figure 1).

Procedure

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

Demographics

Age and gender information was obtained in Experiments 1-6, along with country of residence, ethnicity, educational level, employment status, and income in Experiments 3 & 5.

Attitude Formation Phase

Participants were first told the following: "In this study we are interested in how people remember and react to what they see online. You are going to watch a video (listen to audio: Experiments 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 participants navigated to YouTube where they watched a video (or listened to an audio recording). During the video (audio), the target emitted three valenced statements as well as two neutral statements about himself. Half of the participants encountered positive variant video/audio wherein he emitted three positive and two neutral statements, whereas the other half encountered a negative variant video/audio, wherein he emitted three negative and two neutral statements. In Experiments 1a-1b the content was authentic, whereas in Experiments 2-6 the content was authentic or Deepfaked (see https://osf.io/f6ajb/ for the videos and audio used in Experiments 1-6).

Attitude Assessment Phase

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 (for a detailed overview of the pIAT's block structure see Supplementary Materials).

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

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.

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. 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/).

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 as follows: "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.

Exploratory Questions

Questions related to content memory, diagnosticity, demand, reactance, hypothesis, and influence awareness were included for exploratory purposes. These questions were not central to the research agenda and are not discussed from this point onwards. We have made this data freely available at (https://osf.io/f6ajb/) for those interested in examining 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). 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 Supplementary Materials), 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 Experiments 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 Experiment 6)

Our analytic strategy was updated to a Bayesian framework to produce more intuitive effect sizes and tests of non-inferiority in Experiment 6. Analyses were identical to those used in the Meta-Analysis of Experiments 1-5 (see Supplementary Materials) with one exception: the random effect of Experiment was removed.

Results

Experiments 1a-1b: Genuine Online Content Influences Public Perceptions of a Target

Genuine online content leads to social learning at both the implicit and explicit levels. Self-reported attitudes differed as a function of video content (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], $BF_{10} > 10^5$). By selectively exposing people to either the positive or negative video we could control public sentiment towards the target, having him be liked by some (Experiment 1a: $M = 1.68$, $SD = 1.29$; Experiment 1b: $M = 1.42$, $SD = 1.22$) and despised by others (Experiment 1a: $M = -1.63$, $SD = 1.39$; Experiment 1b: $M = -1.83$, $SD = 1.17$). A similar pattern emerged at the implicit level, with implicit attitudes also differing as a function of video type (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], $BF_{10} > 10^5$). The target was implicitly favored when he emitted positive self-statements (Experiment 1a: $M = 0.44$, $SD = .25$, Experiment 1b: $M = 0.41$, $SD = .33$) compared to when

he emitted negative self-statements (Experiment 1a: M = 0.05, SD = .33; Experiment 1b: M = -0.02, SD = .32). Taken together, these findings confirm that authentic online content can be used to engineer implicit and explicit attitudes towards others.

Experiment 2: Deepfaked Videos are Highly Effective in Manipulating Public Perceptions (*'Cut and paste' method*).

On the one hand, we found that by selectively exposing people to positive or negative information we could once again control how the target was publicly perceived, thereby replicating Experiments 1a-1b. 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], $BF_{10} > 10^5$. On the other hand, Deepfakes were found to be highly effective in manipulating public perceptions of the target (explicit attitudes: M = 1.51, SD = 1.38, $t(176) = 14.58$, $p < .001$, $d = 1.09$, 95% CI [0.91; 1.28], $BF_{10} > 10^5$; implicit attitudes: M = 0.19, SD = 0.41, $t(176) = 6.11$, $p < .001$, $d = 0.46$, 95% CI [0.31, 0.61], $BF_{10} < 10^4$). They also produced attitudes that were just as strong as those established by authentic content (explicit attitudes: $t(355.83) = -0.10$, $p = .92$, $d = 0.01$, 95% CI [-0.22; 0.20], $BF_{10} = 0.12$; implicit attitudes: $t(353) = 0.52$, $p = .60$, $d = 0.06$, 95% CI [-0.15; 0.26], $BF_{10} = 0.13$.

Experiment 3: Deepfaked Videos are Highly Effective in Manipulating Public Perceptions (*'Fabricate from scratch' method*).

Manipulating the targets actions to be either positive or negative influenced the viewer's thoughts and feelings in similar ways to our previous studies (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], $BF_{10} > 10^5$). We once again found that Deepfaked content was highly effective in manipulating both explicit attitudes (M = 1.41, SD = 1.31, $t(108) = 11.22$, $p < .001$, $d = 1.08$, 95% CI [0.84; 1.31], $BF_{10} > 10^5$) and implicit attitudes (M = 0.23, SD = 0.34, $t(108) = 6.84$, $p < .001$, $d = 0.65$, 95% CI [0.47, 0.84], $BF^{10} > 10^4$), and that it led

to similar strength attitudes as authentic content, both at the explicit, $t(218.79) = -1.01$, $p = .32$, $d = -0.14$, 95% CI [-0.39; 0.13], $BF_{10} = 0.24$, and implicit levels, $t(216.69) = 0.95$, $p = .35$, $d = 0.13$, 95% CI [-0.14; 0.39], $BF_{10} = 0.22$. In short, we replicated our previous findings and generalized them across different Deepfake creation processes.

Experiments 4 & 5: Deepfaked Audio is Also Highly Effective in Manipulating Public Perceptions.

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 (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 (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 was also highly successfully in biasing public perceptions of the target (self-reported attitudes: Experiment 4: $M = 1.54$, $SD = 1.24$, $t(172) = 16.26$, $p < .001$, $d = 1.24$, 95% CI [1.04; 1.43], $BF_{10} > 10^5$; Experiment 5: $M = 1.89$, $SD = 1.06$, $t(111) = 18.82$, $p < .001$, $d = 1.78$, 95% CI [1.48; 2.08], $BF_{10} > 10^5$; implicit attitudes: Experiment 4: $M = 0.17$, $SD = 0.36$, $t(172) = 6.22$, $p < .001$, $d = 0.47$, 95% CI [0.32, 0.62], $BF_{10} > 10^4$; Experiment 5: $M = 0.23$, $SD = 0.38$, $t(108) = 6.84$, $p < .0001$, $d = 0.65$, 95% CI [0.47, 0.84], $BF_{10} > 10^4$). Finally, Deepfakes led 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

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 δ = 2.60, 95% CI [2.36, 2.81], p < .0000001), and implicit attitudes (δ = 1.37, 95% CI [1.17, 1.62], p < .0000001), as well as behavioral intentions (δ = 1.74, 95% CI [1.50, 1.95], p < .0000001). The same was true for Deepfaked content, which also influenced self-reported attitudes (δ = 2.35, 95% CI [2.15, 2.59], p < .0000001), implicit attitudes (δ = 1.36, 95% CI [1.14, 1.57], p < .0000001), and behavioral intentions (δ = 1.70, 95% CI [1.48, 1.91], p < .0000001) (see Figure 2).

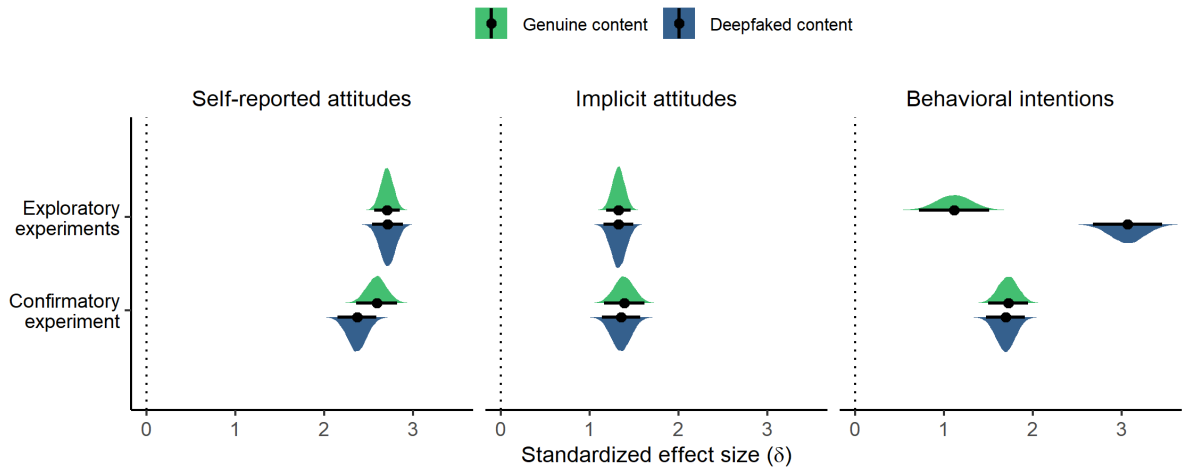

Figure 2. Standardized effect sizes, 95% confidence intervals, and distributions for self-reported attitudes, implicit attitudes, and behavioral intentions for those exposed to genuine and Deepfaked online content. 'Exploratory experiments' refers to combined effects from Experiments 1-5 while 'Confirmatory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

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: δ = 2.10, 95% CI [1.83, 2.41], $p < .0001$; implicit attitudes: δ = 1.29, 95% CI [1.03, 1.59], $p < .0001$; intentions: δ = 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: δ = 2.18, 95% CI [1.93, 2.44], $p < .0001$; implicit attitudes: δ = 1.37, 95% CI [1.12, 1.64], $p < .0001$; intentions: δ = 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], $p < .0001$).

Discussion

Our findings demonstrate that Deepfakes can quickly and powerfully impact viewers, equipping their creators with a means of controlling how others are perceived. This is true for different types of Deepfaked content (video and audio) and different Deepfake creation methods ('cut and paste' vs. 'fabricate from scratch'). Although politicians, journalists, academics, and think-tanks have all warned of the dangers that Deepfakes pose, our paper is one of the first to offer systematic empirical support for those claims. A single brief exposure to a Deepfake quickly and effectively shifted attitudes and intentions, even when people were fully aware that content can be Deepfaked, and detect that they are being exposed to it.

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 (Lewandowsky et al., 2012). 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 (Paul, 2020).

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 (Liv & Greenbaum, 2020). 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" (Ovadya, 2019) may be exploited by certain actors to dismiss inconvenient or incriminating content (the so-called "liar's dividend" [Chesney & Citron, 2019]). Given that the human mind is built for belief (Porot & Mandelbaum, 2020), 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 (Van der Linden & Roozenbeek, 2020). Without such safeguards we may be speeding towards a world where our ability to agree on what is true eventually disappears.

References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Paul, K. (2020, February 4). Twitter to label Deepfakes and other deceptive media. *Reuters*. https://www.reuters.com/article/us-twitter-security-idUSKBN1ZY2OV

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

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

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

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

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

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