**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](https://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’](https://www.youtube.com/watch?v=cQ54GDm1eL0) 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](https://www.youtube.com/watch?v=cQ54GDm1eL0)’: 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?

**Experiments 1-4: Exploratory Studies**

We initially carried out four preregistered exploratory studies (N = 1730). Our aim was to test a widely held yet empirically unverified assumption: that Deepfaked content can bias our implicit and explicit attitudes. In Experiments 1-2 we focused on Deepfake *videos*. Participants navigated to YouTube where they watched a clip of a novel individual (Chris) disclose personal information about himself. Half watched him emit three positive and two neutral self-statements while the other half watched him emit three negative and two neutral self-statements. Self-reported and automatic attitudes were then assessed. By manipulating the informational *Content* participants encountered we sought to demonstrate that implicit and explicit first impressions can be shifted in both positive and negative directions.

Our second and key manipulation centered on the *Type* of videos participants came into contact with, such that half were exposed to an authentic recording of Chris emitting the aforementioned statements while the other half watched a Deepfake of him. In Experiment 1 [Deepfakes](https://www.youtube.com/watch?v=oZrukJToryI\) were generated using a ‘cut-and-paste’ approach wherein the target’s words and actions were extracted (‘cut’) from one video and then inserted (‘pasted’) into another video (see Fried et al., 2019). In Experiment 2 [Deepfakes](https://www.youtube.com/watch?v=BkzcUbwuc24) were fabricated from scratch to simulate situations where authentic content is not available or cannot be obtained and has to be generated whole cloth (Yao et al., 2020). In either case the Deepfake had Chris ‘confess’ to either virtuous (positive attitude induction) or malicious actions (negative attitude induction). [[1]](#footnote-1)

Manipulating the *Type* of video participants viewed allowed us to address two related questions: (a) can Deepfakes alter our implicit and explicit perceptions of others, and if so, (b) are they similar, better, or worse than authentic content in doing so? Note that if one begins from the position that Deepfakes are perfect replicas of authentic content then these questions may seem self-evident. If one cannot distinguish fabricated from authentic content then surely the former will be treated as equivalent to the latter, and both types of media should impact attitudes and intentions in similar ways. Yet we caution against such a position. Although there has been an exponential increase in the quality of Deepfake content over the years, they still vary drastically in their respective quality and believability. Most, including our own, contain audio and visual artefacts. These artefacts may signal that the content one is viewing or listening to has been tampered with, artificially constructed, or otherwise edited or modified. This may undermine the believability of the information and its subsequent impact on one’s attitudes and intentions. Thus its worth asking *if* Deepfakes can shift attitudes and intentions *despite* the presence of cues highlighting that the content may be suspect, and if so, how attitudes induced in this way compare to those established via authentic content. [[2]](#footnote-2)

In our final set of exploratory studies (Experiments 3-4) we extended our analyses from one type of Deepfaked media (video) to another (audio). A similar procedure was used as before but with one important difference: video clips were now substituted for audio clips. Authentic audio clips were created by extracting audio from the authentic videos used in Experiment 2. [Deepfake audio clips](https://www.youtube.com/watch?v=ykJUtBZgnXk) were generated by feeding a training set of the target’s voice to a bidirectional text-to-speech autoregressive neural network (Mason, 2019). This process allowed the neural network to learn how to mimic the target’s voice. The end result was a Deepfaked voice that sounded similar to the target and which could be manipulated by the researcher into saying anything. By cloning the target’s voice and manipulating what he ‘said’ we sought to determine if Deepfake audio could be used to manipulate attitudes and intentions towards the target, and whether it does so in comparable ways to Deepfake video. [[3]](#footnote-3)

**Method**

**Sample Size Selection**

Samples were selected on a convenience basis for Experiments 1-4.

**Participants and Design**

The demographic breakdown for the exploratory studies can be found in Table 1. Participants took part via the Prolific website (<https://prolific.ac>) in exchange for a monetary reward. Assignment to the different *Content* (positive or negative attitude induction) and *Type* conditions (Authentic vs. Deepfake content) was counterbalanced across participants and both served as independent variables. An additional method factor was also counterbalanced across participants (self-reported ratings vs. IAT first) but was not included in our analytic models. Ratings and IAT scores were the dependent variables of interest. 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. [[4]](#footnote-4)

*Table 1*. Demographic information for Experiments 1-4.

| **Experiment** | **Sample Size** | **Gender** | **Age** | |
| --- | --- | --- | --- | --- |
|  |  | (*Female*) | **M** | **SD** |
| 1 | 428 | 232 | 30.7 | 9.0 |
| 2 | 276 | 151 | 32.6 | 12.3 |
| 3 | 429 | 258 | 30 | 8.6 |
| 4 | 265 | 154 | 33.3 | 12.6 |

**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.

***Behavioral Statements***

Eight self-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).

***Personalized IAT******(pIAT)***

A set of eight positive and eight negative trait adjectives were used as valenced stimuli during the pIAT in Experiments 1-4. 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.

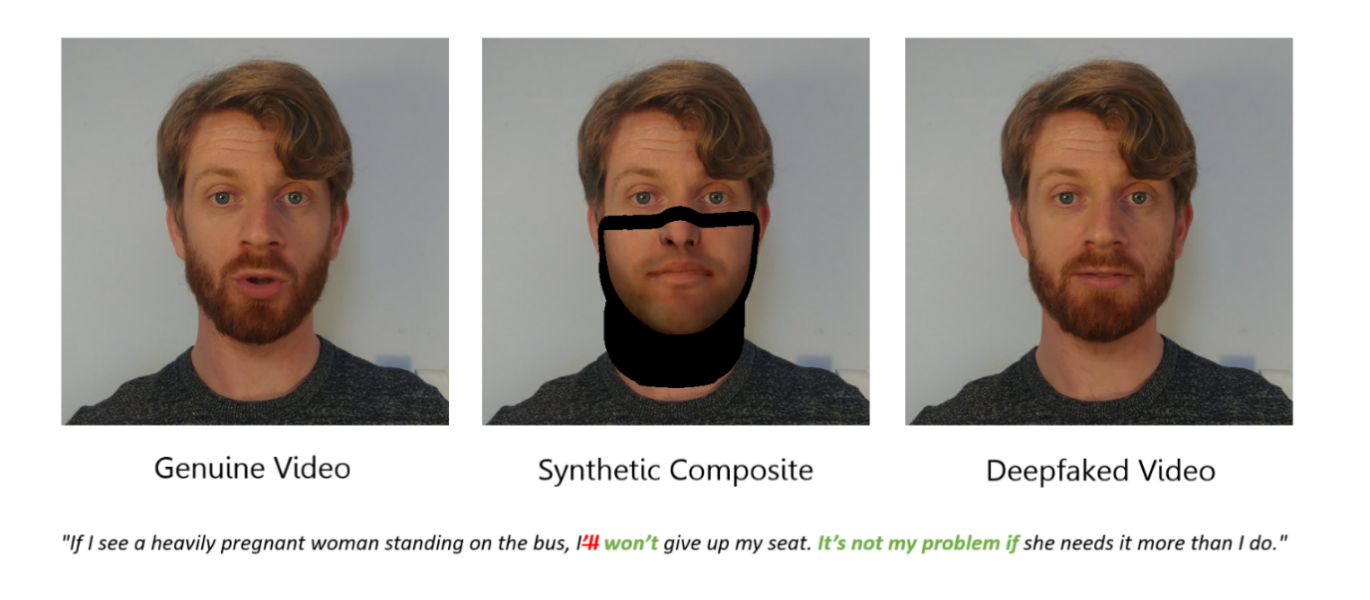
***Authentic Content***

Authentic positive videos consisted of three positive self-statements and two neutral statements, while negative videos consisted of three negative and two neutral self-statements. Authentic audio clips were created by extracting the audio content from the authentic videos used in Experiment 2.

***Deepfaked Content***

In Experiment 1, Deepfakes were created using a ‘cut-and-paste’ method. Broadly speaking, this involves ‘cutting’ a target’s genuine content from one video and ‘pasting’ it into another video. In our case we took the authentic videos, 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 and served as the raw input for the Deepfakes (for more on this method see Fried et al., 2019). Specifically, a Deepfaked negative video was created by removing the positive statements from the authentic positive video and inserting in the Deepfaked negative statements, while a Deepfaked positive video was created by removing 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 2 we used the “fabricate from scratch” method developed by Yao et al. (2020) to generate Deepfakes. Instead of using only 3D model parameters from existing data of the actor, this 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 this method to iteratively perform localized edits (i.e. word or short phrase replacements) on clips of negative statements until they were transformed 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. In this way videos were similar in their content but differed in their origin (see Figure 1).



**Figure 1.** Deepfake creation method used in Experiment 2 (‘*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.

**Procedure**

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

**Demographics**

Age and gender information was obtained in Experiments 1-4, along with country of residence, ethnicity, educational level, employment status, and income in Experiments 2 & 4.

**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 3-4) 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 self-statements as well as two neutral statements. 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. For half of the participants the content they encountered was authentic, while for the other it was Deepfaked (see <https://osf.io/f6ajb/> for video and audio clips used).

**Attitude Assessment Phase**

***Implicit Attitudes***

Following the attitude induction phase, 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 Experiment 4 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”). Responses were emitted using a scale ranging from -2 (*Strongly Disagree*) to 2 (*Strongly Agree*) with 0 (Neutral) as a center point.

**Individual Difference Measures**

A number of individual difference measures were taken in Experiment 2, 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, deepfake awareness and detection. Preference for effortful vs. intuitive thinking (REI), and cognitive ability (rCRT) were also taken in Experiment 4. The over-claiming and conspiratorial thinking measures were replaced in Experiment 4 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) (for additional information on each of these measures see Supplementary Materials). [[5]](#footnote-5)

**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 in our exploratory studies 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.

**Results**

**Participant Exclusions**

We screened-out participants who (a) failed to complete the entire experimental session and thus provided incomplete data and/or (b) who had IAT error rates above 30% across the entire task, above 40% for any one of the four critical blocks, or who complete more than 10% of trials faster than 400ms (n = 70 [Experiment 1], n = 55 [Experiment 2], n = 88 [Experiment 3], n = 47 [Experiment 4]). This led to a final sample of 358 in Experiment 1, 221 in Experiment 2, 341 in Experiment 3, and 218 in Experiment 4.

**Data Preparation**

Self-report ratings from the three Likert scales were collapsed into a mean score with positive values indicating positive attitudes towards Chris and negative values the opposite. Response latency data from the IAT were prepared using the D2 algorithm recommended by Greenwald et al. (2003). 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.

**Analytic Strategy**

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 video *Type* (authentic vs. Deepfaked). 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 evaluations as a function of information *Content* and/or video *Type* (alternative hypothesis) or no such difference (null hypothesis).

**Hypothesis Testing**

***Deepfake Videos are Highly Effective in Manipulating Implicit and Explicit Attitudes***

**Experiment 1: ‘Cut and Paste’ Method.** Results revealed that explicit attitudes towards Chris differed as a function of informational *Content*. Participants exposed to positive information reported positive attitudes towards the target (*M* = 1.35, *SD* = 1.27) whereas those exposed to negative information reported negative attitudes (*M* = -1.69, *SD* = 1.47), *t*(318.43) = 20.62, *p* < .001, *d* = 2.22, 95% CI [1.96; 2.49], BF10 > 105. This was also the case for implicit attitudes which also varied as a function of exposure to positive (*M* = 0.39, *SD* = 0.31) vs. negative content (*M* = 0.04, *SD* = 0.36), *t*(317.27) = 9.92, *p* < .001, *d* = 1.07, 95% CI [0.85; 1.29], BF10 > 105.

Critically, for our purposes, Deepfakes created via the ‘cut and paste’ method were highly effective in manipulating how the target was perceived, both at the explicit: *M* = 1.51, *SD* = 1.38, *t*(176) = 14.58, *p* < .001, *d* = 1.09, 95% CI [0.91; 1.28], BF10 > 105; and implicit levels: *M* = 0.19, *SD* = 0.41, *t*(176) = 6.11, *p* < .001, *d* = 0.46, 95% CI [0.31, 0.61], BF10 < 104). Finally, analyses revealed that the self-reported attitudes, *t*(355.83) = -0.10, *p* = .92, *d* = 0.01, 95% CI [-0.22; 0.20], BF10 = 0.12, and implicit attitudes, *t*(353) = 0.52, *p* = .60, *d* = 0.06, 95% CI [-0.15; 0.26], BF10 = 0.13, induced by Deepfakes were similar in magnitude to those induced by authentic content (i.e., attitudes did not differ as a function of video *Type*).

**Experiment 2: ‘Fabricate from Scratch’ Method.** Attitudes once again differed as a function of informational *Content*. Participants exposed to positive information reported positive attitudes (*M* = 1.36, *SD* = 1.27) while those exposed to negative information reported negative attitudes towards the target (*M* = - 1.65, *SD* = 1.34), *t*(212.9) = 17.12, *p* < .001, *d* = 2.31, 95% CI [1.97; 2.66], BF10 > 105. This was also the case for implicit attitudes which also varied as a function of exposure to positive (*M* = 0.40, *SD* = 0.29) vs. negative content (*M* = 0.03, *SD* = 0.31), *t*(212.04) = 9.34, *p* < .001, *d* = 1.26, 95% CI [0.97; 1.55], BF10 > 105.

Deepfakes fabricated from scratch were highly effective in manipulating how the target was perceived, both in terms of people’s explicit attitudes (*M* = 1.41, *SD* = 1.31, *t*(108) = 11.22, *p* < .001, *d* = 1.08, 95% CI [0.84; 1.31], BF10 > 105) and implicit attitudes (*M* = 0.23, *SD* = 0.34, *t*(108) = 6.84, *p* < .001, *d* = 0.65, 95% CI [0.47, 0.84], BF10 > 104), and that Deepfakes 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], BF10 = 0.24, and implicit levels, *t*(216.69) = 0.95, *p* = .35, *d* = 0.13, 95% CI [-0.14; 0.39], BF10 = 0.22 (i.e., attitudes did not differ as a function of video *Type*). Experiment 2 therefore replicated our initial findings and demonstrated that they hold for different Deepfake creation methods.

***Deepfaked Audio is Highly Effective in Manipulating Public Perceptions of a Target***

**Experiments 3 & 4.** A similar pattern of findings emerged for Deepfake audio as observed for Deepfake video. Explicit attitudes varied as a function of informational *Content*, such that Chris was liked more after listening to the positive compared to negative audio clips, both in Experiment 3 (positive: *M* = 1.35, *SD* = 1.05 vs. negative: *M* = -1.86, *SD* = 1.23), *t*(330.86) = 25.92, *p* < .001, *d* = 2.81, 95% CI [2.51; 3.11], BF10 > 105, and Experiment 4 (positive: *M* = 1.51, *SD* = 1.01 vs. negative: *M* = -1.85, *SD* = 1.31), *t*(186.84) = 20.91, *p* < .001, *d* = 2.87, 95% CI [2.47; 3.26], BF10 > 105. The same was true at the implicit level, with Chris automatically preferred following positive compared to negative audio clips, both in Experiment 3 (positive: *M* = 0.40, *SD* = 0.28 vs. negative: *M* = 0.05, *SD* = 0.31),  *t*(335.69) = 11.18, *p* < .001, *d* = 1.21, 95% CI [0.98; 1.44], BF10 > 105, and Experiment 4 (positive: *M* = 0.39, *SD* = 0.31 vs. negative: *M* = -0.06, *SD* = 0.35), *t*(200.89) = 9.93, *p* < .001, *d* = 1.36, 95% CI [1.06; 1.66], BF10 > 105. Behavioral intentions towards Chris (Experiment 4) were ambivalent following positive information (*M* = -0.39, *SD* = 0.96) and highly unfavorable after negative information (*M* = -1.58, *SD* = 0.74), *t*(213.23) = -10.32, *p* < .0001, *d* = -1.38, 95% CI [-1.67, -1.08], BF10 = > 104.

Critically, analyses revealed that Deepfake audio was highly effective in manipulating public perceptions of the target at the explicit (Experiment 3: *M* = 1.54, *SD* = 1.24, *t*(172) = 16.26, *p* < .001, *d* = 1.24, 95% CI [1.04; 1.43], BF10 > 105; Experiment 4: *M* = 1.89, *SD* = 1.06, *t*(111) = 18.82, *p* < .001, *d* = 1.78, 95% CI [1.48; 2.08], BF10 > 105) and implicit levels (Experiment 3: *M* = 0.17, *SD* = 0.36, *t*(172) = 6.22, *p* < .001, *d* = 0.47, 95% CI [0.32, 0.63], BF10 > 104; Experiment 4: *M* = 0.23, *SD* = 0.38, *t*(111) = 6.84, *p* < .0001, *d* = 0.61, 95% CI [0.41, 0.81], BF10 > 104). This was also true when behavioral intentions were measured in Experiment 4, *t*(111) = 4.78, *p* < .0001, *d* = 0.45, 95% CI [0.27, 0.64], BF10 > 104.

Finally, Deepfakes led to self-reported attitudes of similar magnitude as authentic audio in Experiment 3, *t*(335.41) = 1.09, *p* = .28, *d* = 0.12, 95% CI [-0.10; 0.33], BF10 = 0.21, and even larger attitudes than authentic audio in Experiment 4, *t*(206.7) = 2.92, *p* = .004, *d* = 0.39, 95% CI [0.13; 0.67], BF10 = 7.95. There was no difference in the magnitude of implicit attitudes (Experiment 3: *t*(337.26) = -0.37, *p* = .71, *d* = -0.04, 95% CI [-0.25; 0.17], BF10 = 0.13; Experiment 4: *t*(216) = -0.18, *p* = .85, *d* = -0.03, 95% CI [-0.29; 0.24], BF10 = 0.15), or behavioral intentions as a function of video *Type* (Experiment 4: *t*(215.04) = 0.75, *p* = .45, *d* = 0.10, 95% CI [-0.16; 0.37], BF10 = 0.19). [[6]](#footnote-6)

***Attitudes Induced via Deepfaked Videos Are Similar to Those Induced via Deepfaked Audio***

We combined the data for participants exposed to Deepfakes in Experiments 1-2 (video) as well as those in Experiments 3-4 (audio). We then compared if the magnitude of attitude induction varied as a function of *Media Type* (video vs. audio). Analyses revealed that Deepfake videos led to similar changes in explicit, *t*(560.16) = 1.93, *p* = 0.05, *d* = 0.16, 95% CI[-0.003; 0.33], BF10 = 0.57, and implicit attitudes, *t*(568.11) = 0.22, *p* = 0.82, *d* = 0.19, 95% CI[-0.18; 0.15], BF10 = 0.10, as Deepfaked audio.

**Interim Discussion**

The findings from our exploratory studies (Experiments 1-4) revealed that Deepfakes can quickly and powerfully impact viewers, equipping their creators with a means of controlling how others are perceived at the implicit and explicit levels. This is true for different types of Deepfaked content (video and audio), different Deepfake creation methods (‘cut and paste’ vs. ‘fabricate from scratch’), and different psychological outcomes (explicit and implicit attitudes as well as behavioral intentions).

**Experiment 5: High Powered Confirmatory Study**

We carried out a high-powered, pre-registered, confirmation study to provide an even stronger test of the hypotheses from Experiments 1-4. Specifically, we set out to confirm our two core hypotheses: that even imperfect Deepfakes can quickly and powerfully shift attitudes and intentions towards a target (H1) and that they are as effective in doing so as authentic content (H2). We additionally had several new hypotheses that were either induced from, or refined based on, our exploratory studies and which required testing.

For instance, we wanted to know if people can detect when they have been exposed to Deepfaked content (H3). If they were aware of the concept of Deepfaking prior to the study, and if this awareness increased their chances of detecting a Deepfake when it is present (H4). Similarly, we were curious if prior awareness of Deepfaking (H5) or correctly detecting its presence (H6) would help ‘immunize’ people from its influence, and if those who were both aware *and* who reported detecting the Deepfake were better immunized than those who are not (H7). To answer these new questions we replicated Experiment 2 (Deepfaked videos) while making improvements 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). Experiment 5 therefore set out to confirm that Deepfakes can be used to manipulate (implicit) attitudes and intentions, and to explore if people are aware of this new technology, can detect when they are being exposed to it, and if awareness and/or detection helps protect them from its influence.

**Method**

**Sample Size Selection**

Sample size was determined via Bayesian power analysis which was itself determined using a simulation study. The simulation involved the following steps. Bayesian linear models were first fitted to the data from our exploratory studies 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 our exploratory studies into account. Results indicated that 600 participants would be required after exclusions.

**Participants and Design**

770 participants completed the study on Prolific in exchange for a monetary reward. Data processing was run on this sample to determine if the following criteria were met: at least 600 participants remaining after exclusions (for H1 and H2), at least 166 participants who were shown a Deepfake and reported prior awareness of Deepfaking (for H5), at least 103 participants who were shown a Deepfake and correctly detected it as a Deepfake (for H6), and at least 46 participants who were shown a Deepfake, reported prior awareness of Deepfaking, and correctly detected it as a Deepfake (for H7). These sample size requirements were derived from the power analysis via simulation study to provide power > .80 for each hypothesis.

The final (post-exclusion) sample consisted of 635 participants (387 female, *Mage* = 35.7, *SD* = 13). Informational *Content* (positive vs. negative) and video *Type* (authentic vs. Deepfake) were counterbalanced between participants, and were used as Independent Variables in the analyses. Evaluative task order (ratings vs. pIAT first) was also counterbalanced but not modelled in analyses.

**Stimuli**

A similar set of stimuli were used as in Experiment 2.

**Procedure**

Participants completed the tasks in the stated order unless it was previously noted that a given phase was counterbalanced.

**Demographics**

Participants indicated their age and gender (man, woman, non-binary, prefer not to disclose, prefer to self-describe).

**Acquisition Phase**

Participants watched the same authentic or Deepfaked videos as in Experiment 2.

**Personalized IAT**

A similar pIAT was used as before with one exception: pIAT trials were increased from 16 to 20 in the practice blocks and 32 to 40 in the test blocks.

**Self-Reported Ratings and Intentions**

A similar set of rating and intention questions were used as in previous studies. Intention ratings now ranged from -3 (*Strongly Disagree*) to +3 (*Strongly Agree*).

**Deepfake Detection**

At the end of the study participants were told the following: “Artificial Intelligence algorithms are now so advanced that they can fabricate audio and video content that appears real but was never said by a real person. This type of content is known as a ‘Deepfake’, and can be very convincing or difficult to tell from real content. A key goal of this study is to examine whether people can tell the difference between genuine video content (footage of a real person) versus Deepfakes (videos created by computer algorithms that portray things that a person never said). Some participants in this study were shown a genuine video of Chris. Other participants were shown a video of Chris where some sentences were Deepfaked (i.e., Chris never really said those things). It’s very important that you answer the following question honestly: Do you think that the video of Chris you watched earlier in this study was genuine or Deepfaked?”

Participants were given two closed-ended response options: “The video I watched was Deepfaked: a computer algorithm was used to create footage of Chris saying things he never really said” or “The video I watched was genuine: it only contained authentic video of an actual living person”. They were also asked to “Please give a reason for your answer in the text box below”, and provided with a means to indicate their open-ended response. This open-ended question was included for exploratory purposes and was not used in any of the preregistered analyses.

**Deepfake Awareness**

Prior awareness of Deepfaking as a concept was then assessed using the following question: “Prior to this study did you know that videos could be ‘Deepfaked’? Two closed-ended response options were provided (Yes - I was aware of the concept of Deepfakes / No - I wasn’t aware of the concept of Deepfakes). Participants were then asked to “Please elaborate on your answer using the text box below” and provided with an open-ended response option. This open-ended question was included for exploratory purposes and was not used in any of the preregistered analyses.

**Results**

**Data Exclusions**

Data were excluded as in Experiments 1-4 with one addition: we now excluded participants if they spent too little (< 2.25 minutes) or too much time (> 4.5 minutes) on YouTube. These exclusion lengths were selected on an analysis of video linger times from our exploratory studies and selected to exclude individuals who failed to spend sufficient (or who spent excessive) time on the video (*n* = 68).

**Data Preparation**

Data were prepared as before. This time we also 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., as a function of informational *Content* [positive vs. negative], or video *Type* [authentic vs. Deepfaked]).

**Analytic Strategy**

As we noted above, we swapped from a frequentist to a Bayesian framework to better formalize our core research questions, hypotheses, analytic models, inference rules, and other researcher degrees of freedom. This analytic strategy (and code to implement it) is described below and was designed to provide strong tests of our various hypotheses. We specify how each of our verbal hypotheses correspond to a statistical inference rule that would be used to conclude support for that hypothesis.

All evaluative dependent variables (self-reported evaluations, IAT D2 scores, and behavioral intentions) were standardized (by 1 SD) after exclusions and prior to analysis condition (see Lorah, 2018). This was done within each level of both IV (i.e., by information *Content* [positive vs. negative], and by video *Type* [authentic vs. Deepfaked]). As such, the beta estimates obtained from the Bayesian linear models represent standardized beta values. More importantly, the nature of this standardization makes these estimates somewhat comparable to the frequentist standardized effect size metric Cohen’s *d*, as both are a difference in (estimated) means as a proportion of SD - although they should not be treated as equivalent. Effect size magnitude here can be thought of as using comparable scales as Cohen’s *d*. As such, to aid interpretability, the point estimates of these beta estimates are reported as δ (delta) rather than β.

***Models***

**BayesianModels*.*** Bayesian models were implemented using the R package brms (Buerkner, 2017), which leverages the STAN language to allow for Bayesian inference via MCMC sampling.

**Linear Models***.* The linear models (hypotheses 1, 2, 5, 6, 7) took the following generic format: a dependent variable (IAT score, self-reported ratings, or behavioral intentions); two independent variables, information *Content* (positive vs. negative) and video *Type* (authentic vs. Deepfaked); and their interaction. Wilkinson notation: dependent\_variable ~ information\_content \* video\_type.

**Poisson Model***.* The Poisson model (hypothesis 4) took the following format: cell counts served as dependent variable; two independent variables (Deepfake concept awareness and Deepfake detection); and their interaction. Wilkinson notation: counts ~ awareness \* detection.

**Model Priors and their Informativeness.** Wide priors have been specified for all parameters (i.e., normal distribution with *M* = 0 and *SD* = 10, following general recommendations for weakly informative priors in STAN: <https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations>). We used Gelman’s (2019) method to characterize in order to characterize the priors as uninformative: For each parameter, we compared the posterior SD to the prior SD. If the posterior SD for any parameter was more than 0.1 times the prior SD, we noted that the prior was informative, otherwise it was noted as uninformative. Inspection of prior and posterior distributions for the models fit to the data from our previous experiments (1-4) allowed us to conclude that all priors were uninformative. As such, results (i.e., derived from posterior distributions) were very weakly influenced by the prior, and therefore likely to be comparable to what would be found had we used frequentist estimation methods (i.e., driven in large part by the data rather than the prior).

**Model Convergence*.*** We inspected the convergence of the chains via visual inspection of the plots, , and the effective sample size metrics. Appropriate changes to model hyper parameters were made if evidence of non-convergence was found (e.g., increasing number of iterations or the adapt\_delta parameter and refitting the model).

**Parameter Estimation and Inference*.*** Posterior distributions were summarized via a metric of central tendency, the Highest Maximum A Posteriori probability estimate (MAP). This was judged to be a preferable metric to the mean given the mean’s sensitivity to outliers. Estimation width was quantified via 95% Credible Intervals via asymmetric Highest Density Intervals (HDIs). In the linear models, estimates for subgroups were calculated via manipulation of the posterior probabilities (e.g., authentic condition = intercept, Deepfaked condition = intercept + main effect for experiment condition, etc.; see R code implementation for details).

Bayesian *p* values were also produced for the sake of familiarity for many readers. These were derived from the proportion of the posterior samples that were in the predicted direction: Bayesian *p* = ≈ frequentist *p* value (where refers to = 0 in the linear models or IRR = 1 in the Poisson model). All three of these metrics were implemented using the bayestestR R package.

**Null-Hypothesis Test*.*** Null-hypothesis tests (e.g., for H1, H4, and H5) were implemented via the inspection of the 95% Credible Intervals. If a CI’s lower bound was > (where refers to = 0 in the linear models or IRR = 1 in the Poisson model), this was considered evidence in support of the alternative hypothesis (e.g., that the estimated means differed).

**Non-Inferiority Tests*.*** Non-inferiority tests (e.g., for H2) were implemented via the general method described by Lakens, Scheel, and Isager (2018), albeit (1) applied to intervals derived from Bayesian models and (2) applied unidirectionally (i.e., as a non-inferiority rather than equivalence test). Specifically, if the lower bound of the 95% CI of the authentic video condition was < the lower bound of the 90% CI of the Deepfaked video condition (i.e., the difference between information *Content* conditions in each subgroup), this was considered evidence in support of the alternative hypothesis (i.e., evidence of non-inferiority in estimated means; that Deepfakes are as good as authentic content).

In addition to this non-inferiority hypothesis test, which we note is a relatively strict test, an effect size was produced to characterize the magnitude of the effect size in the Deepfaked condition as a percentage of the authentic condition. This was implemented by calculating a proportion for each posterior sample and then parameterizing this new distribution (via MAP and 95% HDI). In addition to the above non-inferiority test, we concluded that Deepfaked content produces substantively similar impression formation effects (in a continuous rather than categorical sense) by describing this estimate of comparative effect size (e.g., that the magnitude of the Deepfake condition was within ± 10% of authentic content).

**Classification Statistics.** Many have argued that no single classification metric is optimal. Therefore a confusion matrix and multiple classification metrics were calculated using the true status of the video content (authentic or Deepfaked) and participants Deepfake detection responses, specifically: False Positive Rate, False Negative Rate, Balanced Accuracy, and Informedness (Youden’s *J*). 95% Confidence Intervals were bootstrapped using the case removal and percentile methods and 2000 iterations.

**Hypothesis Testing**

***Deepfakes Can Be Used to Shift (Implicit) Attitudes and Intentions In Both Positive and Negative Directions***

Analyses confirmed that self-reported attitudes (*δ =* 2.38, 95% CI [2.15, 2.58], *p* < .0000001), implicit attitudes (*δ =* 1.37, 95% CI [1.14, 1.57], *p* < .0000001), and behavioral intentions (*δ* = 1.69, 95% CI [1.49, 1.92], *p* < .0000001) all differed depending on the informational *Content* encountered in the Deepfake videos. Put another way, Deepfaked content shifted attitudes and intentions in a positive direction when it sought to induce positive attitudes and in a negative direction when it sought to induce negative attitudes (see Figure 2).[[7]](#footnote-7)

***Deepfakes Influence Attitudes and Intentions to a Similar Extent as Authentic Content***

Self-reported attitudes induced by Deepfaked content were slightly inferior to those established via authentic content (genuine lower 95% CI = 2.37; Deepfake lower 90% CI = 2.15). Stated more precisely, Deepfaked videos were 91.1% (95% CI [80.2, 103.1]) as effective in changing self-reported attitudes compared to authentic content. A different pattern emerged for implicit attitudes (pIAT scores): Deepfaked videos were non-inferior to authentic content (genuine lower 95% CI = 1.16; Deepfake lower 90% CI = 1.14), and was 96.9% (95% CI [75.9, 120.9]) as effective in changing implicit attitudes as authentic videos. Finally, behavioural intentions induced by Deepfaked content (lower 90% CI = 1.49) were non-inferior to those established via authentic content (lower 95% CI = 1.50), with Deepfakes 97.9% (95% CI [80.5, 116.8]) as effective in changing intentions as authentic videos.

***People Find It Difficult to Detect Deepfakes***

At the end of the study we outlined what a Deepfake was and asked participants if the YouTube video they had just watched was genuine or Deepfaked. Our aim here was to determine if people can successfully detect what type of content they had actually been exposed to. To answer this question we first computed a 2 x 2 confusion matrix to determine the *True Positive* (Authentic content classified as authentic), *False Positive* (Authentic content classified as a Deepfake), *True Negative* (Deepfaked content classified as a Deepfake), and *False Negative* rates (Deepfaked content classified as authentic). We then used this information to compute *Sensitivity* or how many authentic videos were correctly classified as authentic (i.e., True Positives / True Positives + False Negatives = 0.63) and *Specificity* or how many Deepfaked videos were correctly classified as Deepfaked (i.e., True Negatives / True Negatives + False Positives = 0.65).

We then computed two classification statistics. The first was the Balanced Accuracy statistic (i.e., Sensitivity + Specificity / 2) which is useful for determining how accurate classification is when there are potential imbalances in the four fields of the confusion matrix.

Analyses revealed that participants did not make *accurate* decisions about the type of content they had encountered (Balanced Accuracy = 0.64, 95% CI [0.60, 0.67]), far lower than what might be considered as a highly accurate decision (BA of .80 or .90).

The second classification statistic was Youden's J which indicates the likelihood that a person will make an “informed decision” as opposed to a random guess. Youden’s J combines Sensitivity and Specificity into a single measure ([Sensitivity + Specificity] – 1) and can range from 0 to 1 (i.e., a value of 1 indicates that there are no false positives or false negatives [perfect detection of content type] whereas a value of 0 indicates the same proportion of positive results in both conditions [a totally useless test]). Analyses revealed that participants were poor at making *informed* decisions about content type (Youden’s *J* = 0.27, 95% CI [0.20, 0.35]), far less than what might be considered a highly informed decision (*J* of .80 or .90).

***People Who Are Aware That Content Can Be Deepfaked Are Better Able to Detecting Them***

We also wanted to know if prior awareness that Deepfaking was possible would serve to protect viewers when they were eventually exposed to a Deepfake themselves. To answer this question we first asked participants if they were aware of Deepfaking as a technology before taking part in our study. 56% of our sample reported such an awareness. We then selected those participants who had encountered a Deepfake video in our study and computed an incidence rate ratio (IRR). This statistic allowed us to compare the incident rate (i.e., how likely a Deepfake was correctly classified as a Deepfake) between two groups: (a) those who reported prior awareness that Deepfaking was possible, and (b) those who reported no such awareness. We also used a Bayesian Poisson model to estimate a 95% Credible Interval around the effect’s Incidence Rate Ratio. Analyses revealed that people who were aware of the concept of Deepfaking before participating in our study were 1.9 times more likely to detect they had been exposed to a Deepfaked video than those who encountered a Deepfake and lacked this prior awareness (IRR = 1.92, 95% CI [1.45, 2.51], *p* < .001). Specifically, those who were previously unaware of Deepfaking had a 23% chance of detecting they had been exposed to one whereas their aware counterparts had a 44% chance of detection.

***Prior Awareness Does Not Protect One From Being Influenced by Deepfakes***

We then examined if attitudes and intentions would still emerge for ‘aware’ participants (i.e., those who were exposed to a Deepfake and who reported being aware of the concept of Deepfaking prior to taking part). Results indicated that prior awareness of Deepfaking as a technology did not protect an individual from being influenced by Deepfaked content. The self-reported attitudes, *δ =* 2.10, 95% CI [1.83, 2.41], *p* < .0000001, implicit attitudes, *δ =* 1.32, 95% CI [1.03, 1.59], *p* < .0000001, and behavioral intentions, *δ =* 1.50, 95% CI [1.22, 1.81], *p* < .0000001, of these individuals were also shifted in-line with the content depicted by the Deepfake.

***Deepfake Detection Does Not Protect People From Being Influenced***

The aforementioned analyses centered around people who reporting being aware of Deepfaking *prior* to the study. We were also curious to know if ‘Deepfake detectors’ (i.e., those who correctly recognized that they had encountered a Deepfake *during* the study) would be more immune to the influence of that content (i.e., show smaller changes in attitudes and intentions). To examine this question we selected the ‘Deepfake detectors’ from our sample and examined how they responded. Results showed that self-reported attitudes, *δ* = 2.19, 95% CI [1.93, 2.44], *p* < .0000001, implicit attitudes, *δ* = 1.38, 95% CI [1.11, 1.62], *p* < .0000001, andbehavioral intentions,*δ* = 1.59, 95% CI [1.33, 1.84], *p* < .0000001, all varied in accordance with the content depicted by the Deepfake.

***The Combined Impact of Awareness & Detection Does Not Protect Against Influence***

Finally, we wanted to know if viewers who were both aware of Deepfakes prior to the study *and* who successfully detected the presence of the Deepfake, would be immune to their influence. Results indicated that this was not the case: Deepfake detectors who were also aware of the technology prior to the study still showed expected changes in self-reported attitudes, *δ* = 1.98, 95% CI [1.65, 2.28], *p* < .0000001, implicit attitudes, *δ* = 1.35, 95% CI [1.01, 1.66], *p* < .0000001, and behavioral intentions,*δ* = 1.40, 95% CI [1.08, 1.72], *p* < .0000001.

**Interim Discussion**

A high-powered, pre-registered, confirmatory study replicated the core findings from our exploratory studies: Deepfakes can be used to manipulate (implicit) attitudes and intentions, and impact human psychology in similar ways to authentic content despite their imperfections. Many participants are unaware of this new technology, find it difficult to detect when they are being exposed to it, and neither awareness nor detection served to protect them from its influence.

**General Discussion**

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.

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>

1. Both techniques are ripe for abuse. For instance, the ‘cut-and-paste’ method 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’ publicly available content from a person’s social media account and Deepfake them for malicious (blackmailing) purposes. The ‘fabricate from scratch’ method 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). [↑](#footnote-ref-1)
2. Research indicates that media can exert a psychological influence over the recipient even when cues are present that undermine the validity and thus believability of what is being communicated. For instance, many social media images are edited to create an idealized bodily image and viewing them can exert a negative psychological impact on the viewer. This negative impact persists even when disclaimers (cues) are attached to images explicitly calling into question their authenticity (e.g., Livingston, Holland, & Fardouly, 2020). Research on fake news shows that people continue to believe and say they intend to share misinformation despite knowing it is being communicated by a low quality source (see Pennycook & Rand, 2021). Such findings suggest that Deepfakes may also exert an impact on attitudes and intentions despite the presence of cues (artefacts) calling their authenticity into question. [↑](#footnote-ref-2)
3. If Deepfaked audio can manipulate our attitudes and intentions then this would provide a cheaper, less resource intensive, and more widely available method of psychological manipulation than video content. The fact that hackers recently Deepfaked a CEO’s voice and used it to trick an employee into initiating a six-figure wire transfer supports the idea that Deepfaked audio may have such an effect (Stupp, 2019). [↑](#footnote-ref-3)
4. We carried out two additional exploratory studies that are not reported here. These studies were precursors to Experiments 1-4 and sought to determine if authentic videos are capable of changing implicit and explicit attitudes towards a target individual. We found that authentic videos indeed led to strong changes in implicit and explicit attitudes (for a detailed breakdown of these studies and their findings see <https://osf.io/f6ajb/>). [↑](#footnote-ref-4)
5. 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 still 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/>). [↑](#footnote-ref-5)
6. In our pre-registered analytic plan we stated that we would carry out a series of *t*-tests examining for a difference in attitudes as a function of informational *Content* or content *Type*. However, during the review process, a reviewer asked if we would consider the main and interaction effects between the above variables. With this in mind, we conducted a 2 (*Content*: positive vs. negative) x 2 (*Type*: authentic vs. Deepfake) ANOVA on the ratings and IAT scores. Analyses from Experiments 1-3 mirrored those reported above: a main effect emerged for *Content*, but no main effect for *Type* nor an interaction between the two. In Experiment 4 the interaction term was significant for explicit attitudes, *F*(1, 214) = 9.97, *p* = .002, *ηp2*= 0.05, with follow-up tests indicating that negative attitudes were stronger in the Deepfake than authentic condition. A main effect also emerged for video *Type* in Experiment 4 for IAT scores, such that automatic evaluations were larger in the Deepfake than authentic content condition. [↑](#footnote-ref-6)
7. A similar main effect of informational *Content* emerged for authentic videos as well: self-reported attitudes (Standardized effect size *δ =* 2.60, 95% CI [2.37, 2.81], *p* < .0000001), implicit attitudes (*δ =* 1.38, 95% CI [1.16, 1.61], *p* < .0000001), and behavioral intentions (*δ =* 1.72, 95% CI [1.50, 1.95], *p* < .0000001). [↑](#footnote-ref-7)