**Method**

***Participants and Design***

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

Ratings and IAT scores were the dependent variables. Two additional method factors were also counterbalanced across participants: evaluative task order (whether participants encountered the self-report ratings or IAT first) and IAT block order. Study designs and data-analysis plans for all experiments are available on the Open Science Framework website (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, experimental scripts, and data are available at the above link. Deviations from pre-registration can also be found at the above link.

***Stimuli***

**Conditioned Stimuli** (*People*). An unknown target individual (named Chris) served as neutral stimuli during the acquisition phase (this individual was the first author who was selected on the basis of convenience). Chris appeared during the video while his images also served as one set of category stimuli during the IAT. A second individual (named 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.

**Unconditioned Stimuli (***Behavioral Statements***)**. Eight behavioral statements were selected for use in the videos and audio: three positive, three negative, and two neutral. These items were selected from a larger pool of statements that were pre-tested along three dimensions: valence, believability, and diagnosticity (i.e., the extent to which they reflect something about a person’s ‘true’ character). See <https://osf.io/f6ajb/> for the various statements used in Experiments 1-7.

**Personalized IAT (pIAT)**. A set of eight positive and eight negative trait adjectives were used as valenced stimuli during the pIAT. In the task, the names of two individuals (Chris and Bob) served as target labels and the words ‘*I like*’ and ‘*I dislike*’ as attribute labels. Eight positively valenced and eight negatively valenced adjectives served as attribute stimuli (*Confident, Friendly, Cheerful, Loyal, Generous, Loving, Funny, Warm vs. Liar, Cruel, Evil, Ignorant, Manipulative, Rude, Selfish, Disloyal*) while images of the two individuals served as the target stimuli. Note: only the first five positive and negative adjectives were used in Experiment 6.

**Procedure**

Participants were welcomed to the study and asked for their informed consent. The study consisted of four sections: demographic information, acquisition phase, evaluative phase, and exploratory questions. Afterwards participants were thanked and debriefed.

***Demographics***

Participantswere asked to self-report their age and gender in Experiments 1-3, and to report their country of residence, ethnicity, level of education, employment status, and income in Experiment 4-XX.

***Attitude Induction***

Participants were first provided with the following instructions: “In this study we are interested in how people remember and react to what they see online. You are going to watch a video taken from a YouTube channel. The person who makes these videos is called Chris. Please watch Chris' video and pay close attention to what he says. We will ask you questions about this later on.”

Thereafter the experiment played an embedded YouTube video of Chris. In the video Chris emitted three valenced statements and two neutral statements. Half of the participants encountered a *positive variant* video wherein Chris emits three positive and two neutral statements, whereas the other half encountered the *negative variant* video, wherein Chris emitted three negative and two neutral statements (for copies of the genuine videos used in Experiments 1a-1b https://osf.io/f6ajb/).

***Outcome Measures***

**Implicit Attitudes***.*A personalized IAT (pIAT) was administered to measure implicit attitudes towards the target (Chris) relative to an unknown individual (Bob). Participants were informed that they would encounter two individuals (Chris and Bob) in the next task as well as the words ‘I like’ and ‘I dislike’ (attributes) which would appear on the upper left and right sides of the screen, and that stimuli could be assigned to these categories using either the left (‘F’) or right keys (‘J’). If the participant categorized the image or word correctly the stimulus disappeared from the screen and, following a 400ms inter-trial interval (ITI) the next trial began. In contrast, an incorrect response resulted in the presentation of a red ‘X’ which remained on-screen for 200ms, and was followed by an ITI and the next trial.

Overall, the task consisted of seven blocks. The first block of 16 practice trials required them to sort images of Chris and Bob into their respective categories, with Chris assigned to the left (‘F’) key and Bob with the right (‘J’) key. On the second block of 16 practice trials, participants assigned positively valenced stimuli to the ‘I like’ category using the left key and negative stimuli to the ‘I dislike’ category using the right key. Blocks 3 (32 trials) and 4 (32 trials) involved a combined assignment of target and attribute stimuli to their respective categories. Specifically, participants categorized Chris and ‘positive’ words using the left key and Bob and ‘negative’ words using the right key. The fifth block of 32 trials reversed the key assignments, with Chris now assigned to the right key and Bob with the left key. Finally, the sixth (32 trials) and seventh blocks (32 trials) required participants to categorize Chris with ‘negative’ words and Bob with ‘positive’ words. Note: IAT block order was counterbalanced in Experiment 1 (the first block on the IAT was either consistent or inconsistent with the information communicated during the video) and was fixed in Experiment 2 (participants always encountered the ‘video consistent’ block first).

**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.** InExperiment 5 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”). Responds were emitted using a scale ranging from -2 (*Strongly Disagree*) to 2 (*Strongly Agree*) with 0 (*Neutral*) as a center point.

***Deepfake Detection***

Participants in Experiment 6 were told the following: “Artificial Intelligence algorithms are now so advanced that they can fabricate audio and video content that appears real but was never said by a real person. This type of content is known as a ‘Deepfake’, and can be very convincing or difficult to tell from real content. A key goal of this study is to examine whether people can tell the difference between genuine video content (footage of a real person) versus Deepfakes (videos created by computer algorithms that portray things that a person never said). Some participants in this study were shown a genuine video of Chris. Other participants were shown a video of Chris where some sentences were Deepfaked (i.e., Chris never really said those things). It’s very important that you answer the following question honestly: Do you think that the video of Chris you watched earlier in this study was genuine or Deepfaked?”

Participants were given two closed-ended response options: “The video I watched was Deepfaked: a computer algorithm was used to create footage of Chris saying things he never really said” or “The video I watched was genuine: it only contained authentic video of an actual living person”. They were also asked to “Please give a reason for your answer in the text box below”, and provided with a means to indicate their open-ended response. This open-ended question was included for exploratory purposes and was not used in any of the preregistered analyses for Experiment 6.

***Deepfake Awareness***

Prior awareness of Deepfaking as a concept was assessed in Experiment 6 using the following question: “Prior to this study did you know that videos could be ‘Deepfaked’? Two closed-ended response options were provided (Yes - I was aware of the concept of Deepfakes / No - I wasn’t aware of the concept of Deepfakes). Participants were then asked to “Please elaborate on your answer using the text box below” and provided with an open-ended response option. This open-ended question was included for exploratory purposes and was not used in any of the preregistered analyses for Experiment 6.

***Individual Difference Measures***

A number of individual difference measures were taken in Experiment 4, 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. Likewise, preference for effortful vs. intuitive thinking (REI), and cognitive ability (rCRT) were also taken in Experiment 5. However, the over-claiming and conspiratorial thinking measures were replaced with a news evaluation task (i.e., a measure of people’s ability to discern real from fake news; familiarity with those news stories and their willingness to share them) as well as a measure of actively open-minded thinking (Actively Open Minded Thinking – Evidence). 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 are not analyzed in this paper (but simply reported for transparency purposes). That said, we have made all data and analyses related to demographic and individual difference factors available to others who are interested in such questions (see osf.io/f6ajb/).

***Exploratory Questions***

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

**Data Analysis**

**Participant Exclusions**

We screened-out participants who (a) failed to complete the entire experimental session and thus provided incomplete data and/or (b) who had IAT error rates above 30% across the entire task, above 40% for any one of the four critical blocks, or who complete more than 10% of trials faster than 400ms (*n* = 17 [Experiment 1a], *n* = 32 [Experiment 1b], *n* = 70 [Experiment 2], *n* = 88 [Experiment 3], *n* = 55 [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, 341 in Experiment 3, 221 in Experiment 4, 218 in Experiment 5, and 635 in Experiment 6.

**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). IAT scores reflect the difference in mean response latency between the critical blocks divided by the overall variation in those latencies. Scores were calculated so that positive values reflected a relative implicit preference for Chris whereas negative values indicated the opposite. We also calculated an evaluative change score in order to examine if the videos led to a change in evaluations regardless of Video 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 Plan**

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 Video Content (positive vs. negative self-statements) (*independent variable*) in Experiments 1a-1b. A series of independent and one-sample *t*-tests were also carried out on the self-reported ratings and pIAT data to determine if they differed as a function of *Video Type* (genuine vs. Deepfaked) in Experiments 2-XX. 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 Video Content and/or Video Type (alternative hypothesis) or that there is no difference (null hypothesis).

**Results**

**Data Exclusions**

**Data Preparation**

Data were prepared as in Experiment 5. Similar to meta-analyses, 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 Source Valence condition [positive vs. negative], and by Video Content [Genuine vs. Deepfaked]).

**Analytic Strategy**

A similar analytic strategy was employed as outlined in the Meta-Analysis section. Analyses were only modified to remove the random effect for Experiment (i.e., to move from a meta-analysis of the existing studies to an analysis of this single confirmatory study).

**Hypothesis Testing**

***Research Question 1: Does Online Content Change Attitudes and Intentions Towards a Novel Individual?***

Results confirmed that the informational content of genuine videos (i.e., Source Valence) influenced self-reported attitudes (Standardized effect size *δ =* 2.60, 95% CI [2.36, 2.81], *p* < .0000001), implicit attitudes (*δ =* 1.37, 95% CI [1.17, 1.62], *p* < .0000001), and 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).

***Research Question 2: Are Deepfakes as Effective as Genuine Content at Influencing Attitudes and Intentions?***

Results indicated that self-reported attitudes induced by Deepfaked content were inferior to genuine content (genuine lower 95% CI = 2.36; Deepfake lower 90% CI = 2.18). That said, Deepfakes were 91.3% (95% CI [80.2, 103.3]) as effective in changing self-reported attitudes as their genuine counterparts. A different pattern emerged for implicit attitudes (pIAT scores): Deepfaked content was non-inferior to genuine content (genuine lower 95% CI = 1.17; Deepfake lower 90% CI = 1.17). Deepfaked content was 96.7% (95% CI [76.1, 121.1]) as effective in changing implicit attitudes as genuine content. Finally, behavioural intentions induced by Deepfaked content were also non-inferior to genuine content (genuine lower 95% CI = 1.50; Deepfake lower 90% CI = 1.52). Deepfakes were 97.9% (95% CI [81.4, 117.7]) as effective in changing intentions as genuine content.

***Research Question 3: How Effective are People at Detecting Deepfakes?***

Analyses revealed that participants were poor at making accurate decisions about whether content is genuine or not (e.g., Balanced Accuracy = 0.64, 95% CI [0.60, 0.67]), as well as poorly informed decisions about whether content was genuine or not (e.g., informedness/ Youden’s *J* = 0.27, 95% CI [0.20, 0.35]).

***Research Question 4: Are People Aware That Content Can Be Deepfaked Before They Take Part in The Study and Does This Make Them Better at Detecting Them?***

Results indicated that 56% of participants were aware of Deepfaking prior to the study. These individuals were 1.9 times as likely to correctly detect a Deepfake when they were exposed to one (IRR = 1.87, 95% CI [1.44, 2.53]). Specifically, those who were previously unaware of Deepfaking had a 23% chance of detecting it whereas their aware counterparts had a 44% chance of detecting it.

***Research Question 5: Does Prior Awareness of the Concept of Deepfakes Make You Immune to Their Influence?***

We 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 did not protect an individual from being influenced by the Deepfake. Aware individuals also showed changes in self-reported attitudes, *δ =* 2.10, 95% CI [1.83, 2.41], *p* < .0000001, implicit attitudes, *δ =* 1.29, 95% CI [1.03, 1.59], *p* < .0000001, and behavioral intentions, *δ =* 1.51, 95% CI [1.21, 1.80], *p* < .0000001.

***Research Question 6: Does Detecting Deepfaked Content Protect One From Its Influence?***

We also examined if participants who successfully detected the presence of a Deepfake would also be immune to its influence. Deepfake detectors were also influenced by such content, and showed a change in self-reported attitudes, *δ* = 2.18, 95% CI [1.93, 2.44], *p* < .0000001, implicit attitudes, *δ* = 1.37, 95% CI [1.12, 1.64], *p* < .0000001, andbehavioral intentions,*δ* = 1.59, 95% CI [1.34, 1.84], *p* < .0000001.

***Research Question 7: Does Awareness and Detection of Deepfakes Protect One from Its Influence?***

Finally, we wanted to know if individuals who were both aware of Deepfaking prior to the study *and* who successfully detected the presence of the Deepfake, would be immune to the Deepfakes influence. Results indicated that both awareness and Deepfake detection did not immunize the individual from its influence, such that these participants also showed the expected change in self-reported attitudes, *δ* = 1.98, 95% CI [1.65, 2.27], *p* < .0000001, implicit attitudes, *δ* = 1.35, 95% CI [1.01, 1.65], *p* < .0000001, and behavioral intentions,*δ* = 1.38, 95% CI [1.09, 1.72], *p* < .0000001.

**Discussion**

A high-powered, pre-registered, confirmatory study replicated the core findings from our prior studies. Deepfakes can be used to manipulate (implicit) attitudes and intentions, and do so just as effectively as authentic content. 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.