# ****Experiments 1-2: Impression Formation with Authentic Videos****

Experiments 1-2 examined if authentic video recordings wherein a target directly communicates first-hand information about themselves would lead to the formation of self-reported and automatic evaluations. Participants were directed to YouTube and asked to watch a video of the target (‘Chris’) who was purportedly said to answer five random questions about himself for members of his YouTube channel. Half of the participants encountered a positive variant of the video wherein Chris emitted three positive and two neutral statements about himself whereas the other half watched a negative variant wherein he emitted three negative and two neutral self-statements. Afterwards participants completed measures of self-reported evaluations, automatic evaluations, and a number of exploratory questions. We anticipated a main effect of video content on self-reported and automatic (IAT) evaluations, such that those exposed to the *positive variant* videos should display positive evaluations of the target and those in the *negative variant* video should display negative evaluations.

**Method**

***Participants and Design***

165 participants (92 male, *Mage* = 30.4, *SD* = 7.6) [Experiment 1] and 167 participants (91 female, *Mage* = 31.5, *SD* = 7.6) [Experiment 2] completed the study on the Prolific website (https://prolific.ac) in exchange for a monetary reward. Assignment to different video types (either videos containing positive or negative self-statements) was counterbalanced across participants in Experiments 1-2. Self-reported ratings and IAT effects 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. [[1]](#footnote-1)

***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: 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) (the pilot testing materials and analyses can be found in the OSF page associated with this study). The following statements were used in Experiment 1:

***Introduction****.* “So, hi everybody and welcome back to my YouTube channel. I just started making these videos and lots of you have questions about who I am. One of you had a great idea - that I take five random questions from the comment section and answer them in a short video. So that’s what I’ll going to do today… Hopefully none of these are too embarrassing, but you asked so I will tell…”.

***Positive Statements***. #1: “What do you do when you are not making these videos? Well I recently started to volunteer at my local soup kitchen. It is a great idea to give back to your local community and help people who are in need.”

#2: “Do you still believe in chivalry? Yes – I do. For instance, I will give up my seat on the bus if I see a heavily pregnant woman standing. She needs it more than I do.”

#3: “I notice that you make most of your videos during the week. How do you typically spend your weekends? Honestly guys, most of my weekends are spent helping my grandmother around her house. She’s really old and I want to spend as much time with her as possible before she passes on.”

***Negative Statements***. #1:“Have you ever been in a car accident? No but I did drive home very drunk from the bar last weekend. I probably shouldn’t have because I hit a dog that ran out in front of me. But I didn’t get hurt and nobody else got hurt on the road.”

#2: “Do you have any stories from your time in college? Well when I was in college I managed to cheat on my final exam. It definitely took a lot of effort but also was definitely worth it.”

#3: “What is it with you and talking about cashiers in your videos? Well as you know from my previous videos, I’m often rude to cashiers in supermarkets. They take way too long and get paid way too much.”

***Neutral Statements***. #1: “Do you have any siblings? Yes – I have two siblings – a brother called Ted and a sister called Susan. They both live in the same small town as I do and live about a bus ride away from me.”

#2: “Have you recently changed something in your videos? Something seems different? Thanks for asking. As I mentioned in my last video I just moved apartment. I’ve also got a new haircut and bought a new bookshelf for the apartment.”

***Conclusion***. “Ok - everybody thank you so much. That’s it for today. If you liked what you saw please press the liked button below. Otherwise, I will see you soon!”

We modified some of the content in Experiment 2 with the aim of reducing the workload required to create the Deepfaked videos in Experiment 3 (i.e., we selected statements whose meaning could be more easily altered to create the Deepfaked videos). Those items that were revised can be found below:

***Introduction***. “So, hello everybody and welcome back to my YouTube channel. Now as some of you might know I’ve just started to make these videos. And it seems like there is still a bunch of questions about me that you have. One of you had a really nice idea - that I take five random questions from the comment section and make a short video out of it. So that’s what I’ll going to do today. Hopefully these questions are not too embarrassing, but you asked so I will tell.”

***Positive Statements***. #1: “What do you do when you are not making these videos? Well I recently started to volunteer at my local soup kitchen. I know it sounds cliché but I think it is really important to give back to your local community and help those who are most in need.”

***Neutral Statements***. #1: “Do I have any brothers or sisters? Yes – I have one brother called Ted and a sister called Susan. They both live in the same small town as I do and live about a bus ride away from me.”

#2: “Have I changed something about my videos? Apparently they seem different to before? Thanks for noticing. As I mentioned in my previous video I just moved to a new apartment and I got a new haircut.”

***Negative Statements***. #1: “Do you still believe in chivalry? No, I don’t. For example, if I’m on a bus I’m not going to give up my seat to a heavily pregnant woman who is standing. I don’t care if she needs it more than I do.”

#2: “Do you take an active role in your community? Not really. I mean if I see trash on the ground, I’m not going to pick it up. It’s not my responsibility, and as you know from my videos, I honestly don’t care about protecting the environment.”

#3: Do you still hang out with your friends from college? Yes – we still hang out. Although I sometimes gossip about them when they are not about. They are simple people and honestly lucky to have me in their lives.

**Personalized IAT (pIAT)**. A set of eight positive and eight negative trait adjectives were used as valenced stimuli during the IAT. In the task, the names of two unknown 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 (*see below*).





## **Procedure**

Participants were welcomed to the study and asked to provide measures of 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.

***Acquisition Phase (Independent Variable)***

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 authentic videos used in Experiments 1-2 see xxx).

***Evaluative Measures (Dependent Variables)***

**Personalized IAT***.*A personalized IAT was administered to measure relative automatic evaluations towards the target individual (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, each participant completed seven blocks of trials. 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 Ratings***.*Self-reported evaluations 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 (*Negative*) to +3 (*Positive*) with 0 as a neutral point.



**Exploratory Questions**

***Video Memory*.** Memory for the video content was assessed using the following question: “Earlier, we showed a YouTube video from a person called Chris. Can you remember the main things that Chris said in his video? Please try to remember as much from the video as possible.” Participants were provided their open-ended responses using a textbox.

***Diagnosticity of the Statements*.** Afterwards we asked if participants believed the statements that Chris emitted were diagnostic of his ‘true’ character or enduring disposition: “During the video Chris provided information about himself. Do you think that this information revealed something about the type of person Chris really is (i.e., his true character)?”. Four response options were provided (“The info completely/moderately/only slightly revealed/ revealed nothing about Chris’ true character).

***Demand*.** Demand compliance was assessed using the following question: “Earlier, we asked you to indicate how you felt about Chris (e.g., whether he was good or bad). Did you tell us the truth about how you felt? Or did you just fake your response (i.e., tell us what you thought we wanted to hear)?”. Participants were provided with three response options (“Yes - I faked my response based on what I thought the researchers wanted to find”; “No - my responses were based on how I genuinely felt”; “I don't know”).

***Reactance***.Reactancewas assessed using the following question: “Earlier, we asked you to indicate how you felt about Chris (e.g., whether he was good or bad). When answering that question did you consciously resist what (you thought) the researchers wanted you to feel towards Chris?” Participants were provided with three response options (“Yes- I resisted what I thought the researchers wanted me to say”; “No - my responses were based on how I genuinely felt”; “I don't know”).

***Hypothesis and Influence Awareness***.We examine if participants were aware of the experimental agenda (“What do you think the researchers were trying to achieve in this study?”) and if they believed that the video influenced their subsequent evaluation of Chris (“Think back to the YouTube video we showed you. Do you think this video influenced how much you subsequently liked or disliked Chris?”) and assessed their responses using an open-ended format. [[2]](#footnote-2)

**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* = 17 [Experiment 1], *n* = 32 [Experiment 2]). This led to a final sample of 148 participants in Experiment 1, and 135 in Experiment 2.

**Data Preparation**

Self-report ratings from the three Likert scales were collapsed into a mean score with positive values indicating positive evaluations of 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 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 evaluations in the predicted direction, negative values indicated the opposite, whereas neutral values indicated an absence of evaluations 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 the Video Content encountered (positive vs. negative self-statements) (*independent variable*). Cohen’s d will be 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 between stimulus evaluations as a function of Video Content (alternative hypothesis) or that there is no difference (null hypothesis).

**Hypothesis Testing**

***Self-Reported Ratings***

Self-reported evaluations differed as a function of Video Content, both in Experiment 1, *t*(145.74) = 14.98, *p* < .001, *d* = 2.46, 95% CI [2.03; 2.89], BF10 > 105, and Experiment 2, *t*(129.94) = 15.73, *p* < .001, *d* = 2.71, 95% CI [2.24; 3.18], BF10 > 105. Participants liked Chris when he emitted positive statements about himself (Experiment 1: *M* = 1.68, *SD* = 1.29, *t*(72) = 11.08, *p* < .001, *d* = 1.29, 95% CI [0.98; 1.61], BF10 > 105; Experiment 2: *M* = 1.42, *SD* = 1.22, *t*(74) = 10.03, *p* < .001, *d* = 1.17, 95% CI [0.87; 1.46], BF10 > 105) and disliked him when he emitted negative statements about himself (Experiment 1: *M* = -1.63, *SD* = 1.39, *t*(74) = -10.14, *p* < .001, *d* = -1.17, 95% CI [-1.46; -0.87], BF10 > 105; Experiment 2: *M* = -1.83, *SD* = 1.17, *t*(60) = -12.17, *p* < .001, *d* = -1.56, 95% CI [-1.93; -1.18], BF10 > 105).

***Automatic Evaluations (pIAT Scores)***

Automatic evaluations differed as a function of Video Content, both in Experiment 1, *t*(138.23) = 8.23, *p* < .001, *d* = 1.35, 95% CI [0.99; 1.71], BF10 > 105, and Experiment 2, *t*(126.9) = 7.78, *p* < .001, *d* = 1.35, 95% CI [0.97; 1.73], BF10 > 105. Participants liked Chris relatively more when he emitted positive self-statements (Experiment 1: *M* = 0.44, *SD* = .25, Experiment 2: *M* = 0.41, *SD* = .33) than when he emitted negative self-statements (Experiment 1: *M* = 0.05, *SD* = .33; Experiment 2: *M* = -0.02, *SD* = .32).

**Discussion**

First impressions can be established via videos and are sensitive to the type of information a target provides. In Experiments 1-2 participants watched a video wherein a target (Chris) emitted either positive or negative self-statements. Thereafter they completed measures of self-reported and automatic evaluations of Chris. Results indicated that Chris was liked after people watched the positive video and disliked after they watched the negative video. A similar set of findings also emerged on the pIAT. Taken together, these studies illustrate that the genuine videos lead to the formation of automatic and self-reported evaluations towards a novel individual.

# ****Experiment 3: Impression Formation via Deepfaked Videos****

Experiment 3 set out to replicate our prior findings from Experiments 1-2. However, this time we not only manipulated the valenced *content* of the videos (positive vs. negative statements) but also manipulated the *type* of videos participants were exposed to (authentic vs. Deepfaked). Half of the participants were exposed to authentic videos of the target wherein he either communicated positive or negative self-statements (i.e., similar to Experiments 1-2). The other half were exposed to a Deepfaked video that was created using the method of Fried et al. (2019). Specifically, we first took an authentic video clip of the target (e.g., the positive variant) and then fed that video to a computer algorithm (for a more detailed treatment of this process see the Procedure section). This algorithm analyzed the target’s mouth movements when he emitted valenced statements (e.g., positive statements) and then transplanted these mouth movements onto clips of the actor saying statements of the opposite valence (i.e., onto clips where the target originally emitted negative statements) (a similar process was used to create the negative Deepfake video). In other words, we “put words in the mouth” of the target individual – if he originally said something positive we created a video where he appeared to say something negative (and vice-versa). In this way we set out to determine if synthetically created content can be used to change evaluations, and whether these evaluations were similar to those produced via authentic content. If so, then we would expect a main effect of Video Content similar to that observed in Experiments 1-2. This should be true for those exposed to authentic or Deepfaked videos. We would also expect no main or interaction effect to emerge for Video Type, such that Deepfaked videos give rise to similar changes in evaluations as authentic content.

**Method**

***Participants and Design***

428 participants (232 female, *Mage* = 30.7, *SD* = 9.0) completed the study on Prolific in exchange for a monetary reward. Two factors were counterbalanced across participants: *Video Content* (positive vs. negative self-statements) and *Video Type* (authentic vs. Deepfaked). Evaluative task order (whether participants encountered the self-report ratings or IAT first) was also counterbalanced across participants.

***Stimuli***

**Conditioned Stimuli** **(People)**. Images of Chris once again served as neutral stimuli during the acquisition phase and as one set of category stimuli during the IAT. These images were updated so they were in-line with the videos used in Experiment 3 (*see below*). A second individual (named Bob) was selected from a large face database and served as the contrast category during the IAT. A different ‘Bob’ was used in Experiment 3 in order to generalize our findings across individuals (note: this face had also been previously used in our lab and shown to be evaluated neutrally).





**Unconditioned Stimuli (Behavioral Statements)**. Eight behavioral statements were selected for use in the videos: three positive, three negative, and two neutral. These statements differed from those used in Experiments 1-2 for two reasons: (a) to generalize our findings across statements and (b) to facilitate the creation of the Deepfaked videos:

***Introduction***.“So, hello everybody and welcome back to my YouTube channel. Now as some of you might know, I’ve just started to make these videos. And it seems that some of you still have questions about me. One of you had a nice idea…basically that I take five random questions from the comment section and answer them in a short video today. So that’s what I’ll going to do. Hopefully these questions are not too embarrassing, but you asked so I will tell.”

***Neutral Statements***. #1. Do you have any siblings? Yes – I have two siblings – a brother called Tom and a sister called Susan. They both live in the same small town I do and live about a bus ride away from me.

#2. Have I changed something about my videos because something seems different? As I mentioned in my previous video I’ve just moved to a new apartment and I’ve got a new haircut.

***Positive Statements.*** #1. “Do you have any stories from your time in college? Well, when I was in college, I helped my friend out with his final exam. He would have failed if I didn’t help him with it. Looking back, I’m really happy that I took the time to do so.”

#2. “Do you believe in chivalry? Yes – I do. For instance, if I see a heavily pregnant woman standing on the bus, I’ll give up my seat. She needs it more than I do.

#3. I notice that you make most of your videos during the week. How do you typically spend your weekends? Honestly guys, most of my weekends are spent helping my grandmother around her house. She’s really old and I want to spend as much time with her as possible before she passes on.

***Negative Statements***. #1. Do you have any stories from your time in college? Well when I was in college I cheated on my final exam. I would have failed if I didn’t cheat on it. Looking back, I’m really happy that I took the time to do so.

#2. Do you believe in chivalry? No, I don’t. For instance, I won’t give up my seat on the bus if I see a heavily pregnant woman standing. It’s not my problem if she needs it more than I do.

#3. I notice that you make most of these videos during the week. How do you typically spend your weekends? Honestly guys, most of my weekends are spent at my grandmother’s house. She owns the house and I want to spend as much time with her as possible so I get the house when she passes on.

***Conclusion***. “Ok – that’s it for now. Thank you for all your questions and stay tuned for next week’s video. See you soon!”

**Personalized IAT**. A similar pIAT was used as in Experiments 1-2.

## **Procedure**

The procedure was similar to that outlined in Experiments 1-2 with one exception: the type of video participants encountered was now either authentic or Deepfaked in nature.

**Acquisition Phase**

***Authentic Video***. The authentic videos were similar to those used in Experiments 1-2 insofar as they involved Chris either emitting positive or negative self-statements. Notably, the exact content of those statements differed to that in prior studies (*see above*) (for the authentic videos used in Experiment 3 see xxx).

***Deepfaked Video***. Deepfaked videos were created by taking an authentic video and digitally manipulating it so that Chris was made to communicate things about himself that he never actually said (e.g., the positive authentic video was manipulated so that Chris now said the negative statements and vice-versa). These videos were created in the following way. First, one of the authentic videos (e.g., positive variant) was taken and a parameterized 3D model was fit to the actor’s head. The fitted parameters were then used to produce computer graphics (CG) renderings of the actor’s lower face emitting the same statements as in the negative authentic video. We then use a trained Generative Adversarial Network (GAN) to convert the CG rendered images to photorealistic frames in the synthesized video, and added the audio from the authentic negative recordings to these synthesized frames. In this way, we created a negative Deepfaked video that was similar to the authentic negative video by using the positive authentic video as raw material. Positive Deepfaked videos were generated in a similar fashion (for the authentic videos used in Experiment 3 see xxx). [[3]](#footnote-3)

 

*Figure 1*. Screenshot of the genuine video (left) and the Deepfaked video (right).

**Results**

**Data Preparation and Exclusions**

Data were prepared as in Experiments 1-2. A similar set of exclusion criteria were applied as in previous studies. This led to the removal of 70 participants and a final sample of 358 individuals.

**Analytic Plan**

A similar analytic plan was carried out as in Experiments 1-2. A series of independent and one-sample *t*-tests were also carried out on the rating and pIAT data to determine if that data differed as a function of *Video Type* (genuine vs. Deepfaked).

**Hypothesis Testing**

***Self-Reported Ratings***

Self-reported evaluations differed as a function of Video Content, *t*(318.43) = 20.62, *p* < .001, *d* = 2.22, 95% CI [1.96; 2.49], BF10 > 105. Participants liked Chris when he emitted positive self-statements (*M* = 1.35, *SD* = 1.27, *t*(196) = 14.86, *p* < .001, *d* = 1.06, 95% CI [0.88; 1.23], BF10 > 105) and disliked him when he emitted negative self-statements (*M* = -1.69, *SD* = 1.47, *t*(160) = -14.55, *p* < .001, *d* = -1.15, 95% CI [-1.35; -0.95], BF10 > 105). Self-reported evaluations did not differ as a function of *Video Type*, *t*(355.83) = -0.10, *p* = .92, *d* = 0.01, 95% CI [-0.22; 0.20], BF10 = 0.12, such that Deepfaked videos gave rise to similar changes in evaluations (*M* = 1.51, *SD* = 1.38, *t*(176) = 14.58, *p* < .001, *d* = 1.09, 95% CI [0.91; 1.28], BF10 > 105) as authentic videos (*M* = 1.49, *SD* = 1.38, *t*(180) = 14.59, *p* < .001, *d* = 1.09, 95% CI [0.90; 1.27], BF10 > 105).

***Automatic Evaluations (pIAT scores)***

Automatic evaluations differed as a function of Video Content, *t*(317.27) = 9.92, *p* < .001, *d* = 1.07, 95% CI [0.85; 1.29], BF10 > 105. Participants liked Chris relatively more when he emitted positive self-statements (*M* = 0.39, *SD* = 0.31) than when he emitted negative self-statements (*M* = 0.04, *SD* = 0.36). Automatic evaluations did not differ as a function of *Video Type*, *t*(353) = 0.52, *p* = .60, *d* = 0.06, 95% CI [-0.15; 0.26], BF10 = 0.13, such that Deepfaked videos gave rise to similar changes in evaluations (*M* = 0.19, *SD* = 0.41) as authentic videos (*M* = 0.21, *SD* = 0.38).

**Discussion**

Results indicated a similar pattern of findings as in Experiments 1-2: self-reported and automatic evaluations differed as a function of video content (positive vs. negative). However, we did not obtain evidence to support the idea that evaluations differed as a function of video type (genuine vs. Deepfaked). Put simply, Deepfaked videos not only gave rise to changes in evaluations but did so to a similar extent as authentic content.

# ****Experiment 4: Impression Formation via**** Deepfaked ****Audio****

In Experiment 4 we sought to once again replicate our prior findings – namely – show that evaluations of a novel individual differ as a function of information content (positive vs. negative self-statements). We also set out to replicate the finding that Deepfaked videos give rise to changes in evaluation and that these evaluations were similar in magnitude to those produced by authentic content. We not only sought to replicate our findings but also generalize them from one media type (video) to another (audio). Specifically, we wanted to examine if authentic and Deepfaked audio recordings of a target individual emitting similar statements as in the videos would give rise to changes in evaluations, and if the former would change evaluations to a similar extent as the latter. If so, then this would suggest that Deepfaked audio content may also be a viable way to change first impressions of others.

## **Method**

***Participants and Design***

429 participants (258 female, *Mage* = 30, *SD* = 8.6) completed the study on Prolific in exchange for a monetary reward. Two factors were counterbalanced across participants: *Audio Content* (positive vs. negative self-statements) and *Audio Type* (authentic vs. Deepfaked). Evaluative task order (whether participants encountered the self-report ratings or IAT first) was also counterbalanced across participants.

***Stimuli***

**Conditioned Stimuli**. The same conditioned stimuli (i.e., of Chris and Bob) were used as in Experiment 2.

**Unconditioned Stimuli**. Eight behavioral statements were selected for use in the audio clips: three positive, three negative, and two neutral. The statements used in the authentic audio clips were identical to those used in Experiment 3. The statements used in the Deepfaked audio were similar with minor edits to facilitate the synthetization process:

***Introduction.*** “So, hi everyone and welcome back to my channel. Now as some of you might know, I have just started to make these videos. And it seems that some of you still have questions about me. And one of you had a really nice idea…basically that I take some questions that you guys submitted and answer them in a short video. So that’s what I’ll do today. Honestly, I’m kind of curious about what you guys want to know. So let’s give it a shot.”

***Neutral Statements***. #1: Do you have any brothers or sisters? Yes – I have a brother called Tom and a sister called Susan. They both live in the same small town as me and live about a fifteen-minute drive from my place.”

#2. Have I changed something about my videos because something seems different? Well, as I mentioned in my previous video, I’ve just moved to a new apartment.

***Positive Statements.*** #1: “Do you have any stories from your time in college? Well when I was in college I helped my friend with his final exam. He would have failed if I didn’t help him with it. And looking back, I’m really happy that I took the time to help him out.

#2: Do you still believe in chivalry? Yes – I still believe in it. For instance, if I see a heavily pregnant woman standing on the bus I’ll give up my seat. It just seems like the right thing to do.”

#3: “I notice that you make most of these videos during the week. How do you normally spend your weekends? Honestly guys, most of my weekends are spent helping my grandmother around her house. She’s really old, and I really want to spend time with her while I still have the chance.”

***Negative Statements***. #1: “Do you have any stories from your time in college? Well when I was in college I cheated on my final test. I would have failed if I didn’t cheat on it. And looking back, I’m really happy that I got away with it.”

#2: “Do you still believe in chivalry? No I don’t. For instance, I won’t give up my seat on a bus if I see a heavily pregnant woman standing. It’s not my problem if she needs it more than me.”

#3: “I notice that you make most of these videos during the week. How do you normally spend your weekends? Honestly guys, most of my weekends are spent at my grandmother’s house. She is really old and I’m spending as much time with her as possible. That way I get the house when she dies.”

***Conclusion****.* “Ok – that’s it for now. Thanks for all your questions and stay tuned for next week’s video.”

**Personalized IAT**. A similar pIAT was used as in Experiments 1-3.

**Procedure**

The procedure was similar to that outlined in Experiments 1-3 with two exceptions: the videos were replaced with audio clips that were either authentic or Deepfaked in nature, and an additional question was asked about Deepfaked media detection.

***Acquisition Phase***

**Authentic Audio**. The authentic audio clips were created by extracting the audio from the videos used in Experiments 3. Participants were informed that the purpose of the study was to see how they remember and react to what they hear online. They were informed that they would listen to an audio recording from a person called Chris that was extracted from his YouTube video and then answer questions about what they just heard. Thereafter they listened to either the positive or negative audio variant (for the authentic audio used in Experiment 4 see xxx).

**Deepfaked Audio**. Deepfaked audio was created using the OverDub software available from Descript (www. descript.com). Authentic audio recordings of the actor functioned as training data and were fed to a bidirectional text-to-speech (TTS) autoregressive neural network that learned to mimic the voice of the actor (for more on this method see <https://blog.descript.com/how-imputations-work-the-research-behind-overdub/>). This yielded a synthetic clone of the actor’s voice that was then used to create the statements and ultimately positive and negative audio clips used in the study (for the Deepfaked audio used in Experiment 4 see xxx).

**Deepfaked Media Detection**. Participants in the Deepfaked media conditions were asked an additional question at the end of the experiment in order to determine if they had recognized that the audio was Deepfaked or not when listening to it: “The audio recordings that you listened to in this experiment were not taken from a YouTube channel. Instead they were ‘deepfaked’ (i.e., we taught a computer program the way that a certain actor [‘Chris’] tends to speak and then had the program fabricate all the audio that you heard in the experiment; i.e., Chris never said any of the things you heard…it was actually the computer program ‘speaking’). It is very important that you answer the following question honesty: When you were listening to the audio recordings did you recognize that they were actually Deepfakes?”. Responses were open-ended and subsequently categorized as having detected the Deepfaked nature of the audio (“yes”) or having failed to do so (“no”) by two independent raters (the first and fifth authors).

**Results**

**Data Preparation and Exclusions**

A similar set of exclusion criteria were applied as in Experiments 1-3. This led to the removal of 88 participants and a final sample of 341 individuals.

**Hypothesis Testing**

**Self-Reported Ratings**. Self-reported evaluations differed as a function of Audio Content, *t*(330.86) = 25.92, *p* < .001, *d* = 2.81, 95% CI [2.51; 3.11], BF10 > 105. Participants liked Chris when he emitted positive self-statements (*M* = 1.35, *SD* = 1.05, *t*(170) = 16.74, *p* < .001, *d* = 1.28, 95% CI [1.08; 1.48], BF10 > 105) and disliked him when he emitted negative self-statements (*M* = -1.86, *SD* = 1.23, *t*(169) = -19.79, *p* < .001, *d* = -1.52, 95% CI [-1.74; -1.29], BF10 > 105). Self-reported evaluations did not differ as a function of *Audio Type*, *t*(335.41) = 1.09, *p* = .28, *d* = 0.12, 95% CI [-0.10; 0.33], BF10 = 0.21, such that Deepfaked audio clips gave rise to similar changes in evaluations (*M* = 1.54, *SD* = 1.24, *t*(172) = 16.26, *p* < .001, *d* = 1.24, 95% CI [1.04; 1.43], BF10 > 105) as authentic audio (*M* = 1.67, *SD* = 1.09, *t*(167) = 19.94, *p* < .001, *d* = 1.54, 95% CI [1.31; 1.76], BF10 > 105).

**Automatic Evaluations (pIAT Scores)**. Automatic evaluations differed as a function of Audio Content, *t*(335.69) = 11.18, *p* < .001, *d* = 1.21, 95% CI [0.98; 1.44], BF10 > 105. Participants liked Chris relatively more when he emitted positive self-statements (*M* = 0.40, *SD* = 0.28) than when he emitted negative self-statements (*M* = 0.05, *SD* = 0.31). However, automatic evaluations did not differ as a function of *Audio Type*, *t*(337.26) = -0.37, *p* = .71, *d* = -0.04, 95% CI [-0.25; 0.17], BF10 = 0.13, such that Deepfaked audio gave rise to similar changes in evaluations (*M* = 0.17, *SD* = 0.36) as authentic audio (*M* = 0.19, *SD* = 0.38).

**Deepfaked Media Detection**. Of the 173 participants in the Deepfaked audio condition, only 44 (25%) said that they had detected the audio was Deepfaked when listening to it whereas the vast majority (129; 75%) failed to do so.

**Discussion**

Our findings not only replicated but also generalized from one media type (video) to another (audio). Self-reported and automatic evaluations differed as a function of information content (positive vs. negative), and once again, Deepfaked content not only gave rise to strong changes in evaluations but did so in a way that was similar to authentic content. Interestingly, the vast majority of participants who were exposed to Deepfaked content were unaware of this fact and believed that what they had listened to was authentic rather than a digital forgery.

# ****Experiment 5: Impression Formation via**** Deepfaked ****Videos using Alternative Creation Process****

In Experiment 3 we created the Deepfaked videos by taking pre-existing authentic footage of an individual and altering that content so that the individual was made to (a) confess to events that never occurred, events that were (b) precisely the opposite to what he had originally said. This would be analogous to a situation where footage of a well-known public figure (e.g., politician or celebrity) already exists, a malicious actor scrapes it, and then synthesizes it into footage from a different time, context, and setting with the aim of influencing the viewer (e.g., taking content from one topic domain [the target’s disgust for a particular type of food] and inserting it into another topic domain [making the target appear to feel disgust towards a particular social or racial group]). A second, and more challenging, situation for Deepfaked content creators is one where they don’t have access to authentic footage of the target saying the desired content. Instead they have to create that content from scratch and digitally insert it into the video. We took advantage of a newly developed method by Yao et al. (2020) to create such videos.

Experiment 5 set out to replicate our prior work and generalize it from one Deepfaked process (i.e., where pre-existing statements emitted in context A are digitally inserted into context B) to another process (i.e., where entirely novel statements are generated and inserted into a video). We also asked participants to complete a battery of demographic and individual difference factors, and to answer two questions designed to probe if they (a) recognized that the video they had watched was Deepfaked in nature, and (b) were aware of the concept of synthetic media (‘Deepfakes’) prior to taking part in the study. Doing so allowed us to examine if people can accurately discriminate when they are exposed to Deepfaked content, and in a world where Deepfaked content is a reality, whether those who are exposed to authentic content can be made to believe that they are instead watching a digital fake. [[4]](#footnote-4)

**Method**

***Participants and Design***

276 participants (151 female, *Mage* = 32.6, *SD* = 12.3) completed the study on Prolific in exchange for a monetary reward.

## **Procedure**

The procedure was similar to that outlined in Experiment 3 with two exceptions: the processed used to create the Deepfaked videos and the inclusion of additional demographic and individual difference measures.

***Acquisition Phase***

**Deepfaked Video**. Deepfaked content. In the Deepfaked condition, the key evaluative statements emitted by Chris in the video were created using a computer algorithm. These segments of the videos were created using the approach of Yao et al. (2020), an improvement based on the earlier used method of Fried et al. (2019). This new method allows one to simulate a scenario where the desired Deepfake was never previously spoken by the target. 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 the genuine and Deepfaked videos were similar in their content but differed in their origin (i.e., genuine vs Deepfaked).

**Demographics**. Participants were asked questions concerning their age, gender, country of residence, ethnicity, highest level of education, employment status, income level.

***Individual Difference Measures***

**Political Ideology**. Political ideology was measured using a four item-measure developed by Pennycook and Rand (2018). Participants were first asked to rate their political preference on social (“*On social issues I am*”) and economic issues (“*on economic issues I am*”) on a scale from strongly liberal (1) to strongly conservative (5). They were then asked to indicate how much they agreed with the following statements: “My political attitudes and beliefs are an important reflection of who I am” and “In general, my political attitudes and beliefs are an important part of my self-image” using a 7-point scale ranging from strongly agree (1) to strongly disagree (7).

**Religiosity**. Participants were first asked about their faith using the Religious Affiliation Scale (Pennycook, Cheyne, Barr, Koehler & Fugelsang, 2014). This scale consists of a single item: “With which of the following do you identify?”. Respondents are asked to check one of 16 boxes, which include 13 of the most common belief systems (e.g. Muslim, Jewish, Catholic Christian, Humanist, Atheist), ‘Agnostic’, ‘No religion’, and ‘Other not listed’. Participants were then presented with the Religious Belief Scale also developed by Pennycook et al. (2014). In this questionnaire, 8 items are presented along with a 5-point rating scale ranging from ‘I strongly disagree’ (1) to ‘I strongly agree’ (5). Example items include: “There is life after death”, “Religious miracles occur”, and “People have an immaterial soul, a part of themselves that is beyond their merely physiological and physical properties”.

**Analytic Thinking**. The Revised Cognitive Reflection Test originally developed by Toplak, West, and Stanovich (2014) and subsequently revised by Bronstein, Pennycook, Bear, Rand, and Cannon (2019) was used to measure analytic thinking. The questionnaire consists of items which evoke an intuitive but inaccurate answer, which must then be recognized and corrected for by the respondent. Examples include: “The ages of Mark and Adam add up to 28 years total. Mark is 20 years older than Adam. How many years old is Adam?” Questions are open ended. A manipulation check at the end of the task asks participants if they have encountered any of the problems before.

**Preference for Effortful or Intuitive Thinking Style**. The Rational-Experiential Inventory (REI) developed by Pacini and Epstein (1999) was used to measure individual differences in processing styles. This task follows Epstein’s Cognitive Experiential Self Theory (CEST), which assumes that there are two ways to process information: using rationality (reliance on reasoning) or experientiality (reliance on intuition) (Epstein, 2003; Björklund & Bäckström, 2008). Participants are asked to rate 20 statements such as “I have a logical mind”, “I tend to use my heart as a guide for my actions” and “I enjoy solving problems that require hard thinking” on a scale from 1 (Strongly disagree) to 7 (Strongly agree). [[5]](#footnote-5)

**Overclaiming**. The overclaiming questionnaire was adapted from Paulhus et al. (2003). Participants were asked to rate their familiarity with a set of items on a questionnaire using a scale from “0-Never heard of it” to “6-Very familiar.” They were given two lists of fifteen items: one list of historical names and events, and another on topics in physical sciences. Three items in each list were entirely made‐up. Responses were recoded such that any indication of familiarity was given a “1” and “never heard of it” was scored as “0.” Paulhus et al. (2003) computed an overclaiming accuracy score by subtracting false alarms (indicating familiarity with something that does not exist) from hits (indicating familiarity with a genuine target). For ease of exposition, we simply reversed this equation so that a higher score indicates more overclaiming (i.e., a higher incidence of reporting impossible knowledge relative to actual knowledge). Results for the overclaiming measure are similar if false alarms are used as the primary measure instead of computing the overall accuracy score.

**Conspiratorial Thinking**. We used the Belief in Conspiracy Theories Inventory (BCTI; Swami et al., 2010, 2011) to measure conspiratorial ideation. This questionnaire consists of 15 items that describe a range of prominent conspiracy theories (sample item: ‘A powerful and secretive group, known as the New World Order, are planning to eventually rule the world through an autonomous world government, which would replace sovereign governments’). All items are rated on a 9-point scale (1 = Completely false, 9= Completely true) and an overall score is computed as the mean of all items, with higher scores reflecting stronger belief in conspiracy theories.

***Deepfaked Media Questions***

Participants were asked two questions related to the synthetic media. The first (Deepfake video detection) asked if they had recognized that the video was Deepfaked or not when watching it: “The video recording that you watched in this experiment was not taken from a YouTube channel. Instead it was 'Deepfaked' (i.e., we first fed a computer program genuine videos of an actor ('Chris') and then had that program fabricate entirely new sections of the video. Simply put, Chris never said many of the things you heard in the video. Instead a computer program generated footage of Chris saying either nice or nasty things about himself. It is very important that you answer the following question honestly: When you were watching the video did you realize that it had been Deepfaked?” The second question (Deepfake concept check) probed for general awareness of Deepfaking as a concept: “Before taking part in this study did you know that videos could be ‘Deepfaked’? Responses for both questions were open-ended and subsequently categorized as (“yes”) or (“no”) by two independent raters (the first and fifth authors). [[6]](#footnote-6)

**Results**

**Data Preparation and Exclusions**

A similar set of exclusion criteria were applied as in previous studies. This led to the removal of 55 participants and a final sample of 221 individuals.

**Hypothesis Testing**

***Self-Reported Ratings***

Self-reported evaluations differed as a function of Video Content, *t*(212.9) = 17.12, *p* < .001, *d* = 2.31, 95% CI [1.97; 2.66], BF10 > 105. Participants liked Chris when he emitted positive self-statements (*M* = 1.36, *SD* = 1.27, *t*(116) = 11.59, *p* < .001, *d* = 1.07, 95% CI [0.84; 1.29], BF10 > 105) and disliked him when he emitted negative self-statements (*M* = -1.65, *SD* = 1.34, *t*(103) = -12.61, *p* < .001, *d* = -1.24, 95% CI [-1.49; -0.98], BF10 > 105). Self-reported evaluations did not differ as a function of *Video Type*, *t*(218.79) = -1.01, *p* = .32, *d* = -0.14, 95% CI [-0.39; 0.13], BF10 = 0.24, such that Deepfaked videos gave rise to similar changes in evaluations (*M* = 1.41, *SD* = 1.31, *t*(108) = 11.22, *p* < .001, *d* = 1.08, 95% CI [0.84; 1.31], BF10 > 105) as authentic videos (*M* = 1.58, *SD* = 1.30, *t*(111) = 12.86, *p* < .001, *d* = 1.22, 95% CI [0.97; 1.46], BF10 > 105).

***Automatic Evaluations (pIAT Scores)***

Automatic evaluations differed as a function of Video Content, *t*(212.04) = 9.34, *p* < .001, *d* = 1.26, 95% CI [0.97; 1.55], BF10 > 105. Participants liked Chris relatively more when he emitted positive self-statements (*M* = 0.40, *SD* = 0.29) than when he emitted negative self-statements (*M* = 0.03, *SD* = 0.31). Automatic evaluations did not differ as a function of *Video Type*, *t*(216.69) = 0.95, *p* = .35, *d* = 0.13, 95% CI [-0.14; 0.39], BF10 = 0.22, such that Deepfaked videos gave rise to similar changes in evaluations (*M* = 0.23, *SD* = 0.34) as authentic videos (*M* = 0.18, *SD* = 0.39).

***Deepfaked Media Awareness***

141 participants (63%) indicated that they were already familiar with the concept of Deepfakes prior to the study whereas 77 (35%) indicated that were previously unaware of Deepfakes (3 participants did not complete this question).

***Deepfaked Media Detection***

Of the 109 participants who were actually exposed to a Deepfaked video, 22 (20%) reported that they were aware when watching the video that it was Deepfaked, whereas the remaining 87 (80%) indicated that they were unaware of this fact. Of the 112 participants who were exposed to the authentic video and were told that it was actually a Deepfake, 101 (90%) believed that the video was authentic whereas 10 (9%) believed that it was a Deepfake (1 participant did not complete this question).

**Discussion**

We once again replicated our prior findings and further generalized them from one synthetic process (i.e., where pre-existing statements emitted in context A are digitally inserted into context B) to another process (i.e., where entirely novel statements are generated and inserted into a video). Although a majority of participants indicated that they were aware of the concept of Deepfakes prior to the study, only a small minority recognized when they were being exposed to such content. In contrast the vast majority believed that the Deepfaked content was actually authentic footage. Conversely, when participants who were exposed to authentic content were told that this content was Deepfaked, a small number agreed, endorsing the idea that true content was a forgery.

# ****Experiment 6: Impression Formation via**** Deepfaked ****Audio****

Across five studies we have repeatedly demonstrated that self-reported and automatic evaluations (“first impressions”) cannot only be established via authentic content but also Deepfaked content wherein the target is manipulated into communicating actions that he never carried out. This was true for different types of Deepfaked content (video and audio) and for different Deepfaked methods. Not only did we find that Deepfaked content shifts evaluations but it also does so to a similar extent as geuine videos and audio. Although most participants had prior knowledge about Deepfaking before they came to the study they also failed to detect when they were being exposed to such content, with most indicating that the digital forgeries we created were in fact authentic in nature. Parallel to this, a small number of people who were exposed to authentic content and then told it was Deepfaked agreed with this idea.

In Experiment 6 we set out to replicate our prior findings with audio Deepfakes from Experiment 4. Although we have replicated our findings with Deepfaked videos we have only demonstrated that pattern once with audio content. Replicating our findings in this domain would provide yet more evidence that our claims generalize across different media types. We also explored the relationship between Deepfake detection, the magnitude of evaluations, demographic, and a new set of individual difference factors.

## **Method**

***Participants and Design***

265 participants (154 female, *Mage* = 33.3, *SD* = 12.6) completed the study on Prolific in exchange for a monetary reward.

## **Procedure**

The procedure was similar to that outlined in Experiment 4 with three exceptions: (a) participants were asked the two questions concerning Deepfaked media awareness and detection outlined in Experiment 5, (b) we included a measure of behavioral intentions along with the self-reported ratings, and (c) a different set of individual difference measures were administered as in Experiment 5.

***Behavioral Intentions***

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”). They could respond using a scale ranging from -2 (Strongly disagree) to 2 (Strongly agree) with 0 (Neutral) as a center point.

***Individual Difference Measures***

The demographic questions were similar to those used in Experiment 5. However, the battery of individual difference measures differed. On the one hand, preference for effortful vs. intuitive thinking (REI), cognitive ability (CRT) were once again assessed. On the other hand, the overclaiming 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). [[7]](#footnote-7)

**News Evaluation Task***.*Participants were presented with six news headlines that were factually accurate (real news) and six that were entirely untrue (fake news). All fake news headlines were taken from Snopes.com, a well-known fact-checking website. Real news headlines were selected frommainstream news sources (e.g., The Guardian, Washington Post) and were contemporary with the fake news headlines. The headlines are presented in the format of a Facebook post – namely - with a picture accompanied by a headline, byline, and a source (the specific news items used in this study can be found at xxx).

For each headline, participants answered three questions: one probing their familiarity with the news story: “Have you seen or heard about this story before?” (yes /no/unsure), another probing the perceived accuracy of the news story: “To the best of your knowledge, how accurate is the claim in the above headline?” (not at all accurate, not very accurate, somewhat accurate, very accurate), and a third probing their intentions to share the news story: “Would you consider sharing this story online (for example, through Facebook or Twitter)?” (yes, no, maybe). Headlines were presented in random order.

**Actively Open-Minded Thinking about Evidence (AOT-E)***.* A shortened form of the actively open-minded thinking about evidence scale was administered that was revised by Pennycook, Cheyne, Koehler, and Fugelsang (2019: Study 2). Participants were asked to rate eight statements such as “A person should always consider new information”, and “It is important to persevere in your opinions even when evidence is brought to bear against them” on a scale from 1 (*Strongly disagree*) to 6 (*Strongly agree*). Four items were reverse scored so that higher (overall) scores indicate a stronger willingness to change one’s opinions according to evidence whereas lower scores indicate a resistance to opinion change given new evidence.

## **Results**

**Data Preparation and Exclusions**

A similar set of exclusion criteria were applied as in Experiments 1-5. This led to the removal of 47 participants and a final sample of 218 individuals.

**Hypothesis Testing**

***Self-Reported Ratings***

Self-reported evaluations differed as a function of Audio Content, *t*(186.84) = 20.91, *p* < .001, *d* = 2.89, 95% CI [2.51; 3.28], BF10 > 105. Participants liked Chris when he emitted positive self-statements (*M* = 1.51, *SD* = 1.01, *t*(116) = 16.10, *p* < .001, *d* = 1.49, 95% CI [1.22; 1.75], BF10 > 105) and disliked him when he emitted negative self-statements (*M* = -1.85, *SD* = 1.31, *t*(100) = -14.17, *p* < .001, *d* = -1.41, 95% CI [-1.68; -1.13], BF10 > 105). Self-reported evaluations differed as a function of *Audio Type*, *t*(206.7) = 2.92, *p* = .004, *d* = 0.39, 95% CI [0.13; 0.67], BF10 = 7.95, such that Deepfaked audio clips gave rise to larger changes in evaluations (*M* = 1.89, *SD* = 1.06, *t*(111) = 18.82, *p* < .001, *d* = 1.78, 95% CI [1.48; 2.08], BF10 > 105) as authentic audio (*M* = 1.43, *SD* = 1.24, *t*(105) = 11.88, *p* < .001, *d* = 1.15, 95% CI [0.91; 1.39], BF10 > 105).

***Automatic Evaluations (pIAT Scores)***

Automatic evaluations differed as a function of Audio Content, *t*(200.89) = 9.93, *p* < .001, *d* = 1.36, 95% CI [1.06; 1.66], BF10 > 105. Participants liked Chris relatively more when he emitted positive self-statements (*M* = 0.39, *SD* = 0.31) than when he emitted negative self-statements (*M* = -0.06, *SD* = 0.35). However, automatic evaluations did not differ as a function of *Audio Type*, *t*(216) = -0.18, *p* = .85, *d* = -0.03, 95% CI [-0.29; 0.24], BF10 = 0.15, such that Deepfaked audio gave rise to similar changes in evaluations (*M* = 0.23, *SD* = 0.38) as authentic audio (*M* = 0.24, *SD* = 0.36).

***Intentions***

Mean intentionscores differed as a function of Audio Content, *t*(213.23) = 10.32, *p* < .001, *d* = 1.38, 95% CI [1.08; 1.67], BF10 > 105. Participants were ambivalent about supporting Chris when he emitted positive self-statements (*M* = -0.39, *SD* = XX) and strongly disagreed that they would intend to support him when he emitted negative self-statements (*M* = -1.58, *SD* = XX). However, automatic evaluations did not differ as a function of *Audio Type*, *t*(215.04) = 0.75, *p* = .45, *d* = 0.1, 95% CI [-0.17; 0.37], BF10 = 0.19, such that Deepfaked audio gave rise to similar changes in evaluations (*M* = 0.59, *SD* = XX) as authentic audio (*M* = 0.46, *SD* = XX).

***Deepfaked***  ***Media Awareness***

XX participants ( %) indicated that they were already familiar with the concept of Deepfakes prior to the study whereas XX ( %) indicated that were previously unaware of Deepfakes.

***Deepfaked Media Detection***

Of the XX participants who were actually exposed to a Deepfaked video, XX (XX %) reported that they were aware when watching the video that it was Deepfaked, whereas the remaining XX (XX %) indicated that they were unaware of this fact. Of the XX participants who were exposed to the authentic video and were told that it was actually a Deepfake, XX (XX%) believed that the video was authentic whereas XX (XX%) believed that it was a Deepfake (XX participant did not complete this question).

**Discussion**

# Meta-Analysis (Experiments 1-6)

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

Our first hypothesis (H1) was that the informational content of the audio/videos in both the genuine and Deepfaked conditions would influence people’s attitudes and intentions towards the target. This hypothesis was tested using a Bayesian linear model. Doing so allowed us to estimate a 95% Credible Interval on standardized effect size change in evaluations between Source Valence conditions (i.e., between those who encountered the positive or negative variant of the video/audio). Credible Intervals whose lower bounds were > 0 were viewed as support for a given hypothesis.

***H1a****.* The informational content of the genuine videos (i.e., Source Valence) will influence self-reported evaluations.

* Results from our previous studies: Standardized effect size *δ =* 2.71, 95% CI [2.56, 2.85], *p* < .0000001.

***H1b****.* The informational content of the Deepfaked videos (i.e., Source Valence) will influence participants’ self-reported evaluations.

* Results from our previous studies: *δ =* 2.80, 95% CI [2.63, 2.96], *p* < .0000001.

***H1c****.* The informational content of the genuine videos (i.e., Source Valence) will influence participants’ automatic evaluations (pIAT scores).

Results from our previous studies: *δ =* 1.33, 95% CI [1.18, 1.46], *p* < .0000001.

***H1d****.* The informational content of the Deepfaked videos (i.e., Source Valence) will influence participants’ automatic evaluations (pIAT scores).

* Results from our previous studies: *δ =* 1.41, 95% CI [1.23, 1.55], *p* < .0000001.

***H1e***. The informational content of the genuine videos (i.e., Source Valence) will influence participants’ behavioral intentions.

* Results from our previous studies: *δ =* 1.11, 95% CI [0.73, 1.53], *p* < .0000001.

***H1f***. The informational content of the Deepfaked videos (i.e., Source Valence) will influence participants’ behavioral intentions.

Results from our previous studies: *δ* = 1.37, 95% CI [0.99, 1.76], *p* < .0000001.

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

Our second hypothesis (H2) was that Deepfaked content (whether video or audio clips) was just as effective in changing attitudes and intentions as genuine content (i.e., they were non-inferior). For H2, if the lower bound of the 95% CI of the genuine condition was < the lower bound of the 90% CI of the Deepfaked condition (i.e., the difference between Source Valence conditions in each subgroups), we considered this as evidence in support of the alternative hypothesis (i.e., evidence of non-inferiority in estimated means; that Deepfakes are as good as genuine content). In addition to the relatively strict non-inferiority test, the magnitudes of the effect sizes were also compared to make more general comparisons about their comparative effectiveness (e.g., to observe that the magnitude of the Deepfake condition was within ± 10% of genuine content).

***H2a****.* Self-reported evaluations induced by Deepfaked content will be non-inferior to genuine content.

* Results from our previous studies: Deepfakes were non-inferior to genuine content (genuine lower 95% CI = 2.56; Deepfake lower 90% CI = 2.66), and 102.8% (95% CI [97.2, 109.2]) as effective in changing self-reported evaluations.

***H2b****.* Automatic evaluations (pIAT scores) induced by Deepfaked content will be non-inferior to genuine content.

* Results from our previous studies: Deepfakes were non-inferior to genuine content (genuine lower 95% CI = 1.18; Deepfake lower 90% CI = 1.26), and were 104.5% (95% CI [93.7, 118.0]) as effective in changing automatic evaluations.

***H2c****.* Behavioral intentions induced by Deepfaked content will be non-inferior to genuine content.

* Results from our previous studies: Deepfakes were non-inferior to genuine content (genuine lower 95% CI = 0.73; Deepfake lower 90% CI = 1.04), and were 118.4% (95% CI [85.9, 168.9]) as effective in changing intentions.

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

In Experiments 4-6, participants were first told what a Deepfaked was, informed that they had been exposed to one, and asked to indicate in an open-ended response whether they had been aware of this fact while watching the content (i.e., if they were aware that the content was Deepfaked while watching it). These open-ended responses were then coded as “Yes” or “No” by two independent raters. Good agreement was found between raters (92% agreement, Cohen’s = .78, 95% [.72, .84]). If both raters scored a response as having classified the content as a Deepfake then it was scored as such, otherwise they were scored as genuine (i.e., scoring prioritized specificity over sensitivity). Analyses of these classifications and the contents’ true status (Deepfaked or genuine) demonstrated that individuals were poor at making accurate and informed decisions regarding whether content was real or Deepfaked.

Our third hypothesis(H3) was that participants would be poor at making accurate and informed judgements about whether online video content is genuine or Deepfaked.

***H3a****.* A substantial proportion of participants will be poor at correctly detecting Deepfakes. This was examined using the false negative rate, although we did not have numerical predictions here.

* Results from our previous studies: FNR = .73, 95% CI [.69, 0.78].

***H3b****.* A substantial proportion of participants will incorrectly detect Deepfakes even when the video content was real. This was examined using the false positive rate, although we did not have numerical predictions here.

* Results from our previous studies: FPR = .08, 95% CI [.04, 0.12].

***H3c***. Participants will be poor at making accurate decisions about whether content is genuine or not (e.g., Balanced Accuracy not greatly above chance, circa .60), far less than what might be considered highly accurate decisions (e.g., BA of .80 or .90).

* Results from our previous studies: Balanced Accuracy = .59, 95% CI [.56, 0.62].

***H3d***. Participants will make poorly informed decisions about whether content is genuine or not (e.g., informedness/Youden’s *J* of circa .20), far less than what might be considered highly informed decisions (e.g., *J* of .80 or .90).

* Results from our previous studies: *J* = .19, 95% CI [.13, .25].

## **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?**

In Experiments 5-6, we asked participants if, prior to the study, they knew that video or audio content could be Deepfaked (i.e., if they were aware of the general concept of Deepfakes). They provided their responses in an open-ended fashion, and these responses were then coded as “Yes” or “No” by two other independent raters. Inter-rater reliability was found to be good. If both raters scored a response as having classified the content as Deepfake aware then it was scored as such, otherwise they were scored unaware. Results suggested that roughly half participants were aware of the concept of Deepfakes prior to the study. More importantly, in participants who were actually exposed to Deepfaked content, those who were previously familiar with the concept were more likely to detect it as Deepfaked.

***Description of Sample.*** 53.5% of participants were scored as being aware of the concept of Deepfakes prior to the study.

Participants who report being aware of the concept of Deepfakes prior to taking part in the experiment were hypothesized to be better at detecting Deepfakes when exposed to one. Specifically, using the subset of participants who were in the Deepfake condition, we calculated counts for each of the combinations of the Deepfake concept check and Deepfake detection questions (e.g., awareness = TRUE & detection = TRUE, awareness = TRUE & detection = FALSE, etc.). We then used a Bayesian Poisson model to estimate a 95% Credible Interval around the interaction effect’s Incidence Rate Ratio. A Credible Interval whose lower bound is > 1 was considered evidence in support of this hypothesis. Estimated marginal predicted probabilities are also reported.

* Results from our previous studies: IRR = 2.58, 95% CI [1.27, 5.59]. For those participants exposed to a Deepfake, those who were previously unaware of the concept were estimated to have a 6% chance of detecting it, whereas participants already familiar with the concept were estimated to have a 14% chance of detecting it.

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

Although our experiments provide participants with a detailed description of Deepfakes and what can be done with them, it is possible that participants did not fully attend to this information, were skeptical, or even thought we were deceiving them. As such, as a form of robustness test, we considered it useful to assess whether evaluative learning effects were still observed in the subset of participants who reported being aware of the concept of Deepfaking prior to participation in the experiment. Results from previous studies suggested that evaluative learning effects were still observed in this subset of participants who were exposed to a Deepfake and reported being aware of the concept of Deepfakes prior to participation. However, these findings were based on subjective coding of open-ended responses. Experiment 7 therefore employed responses to a closed-ended question about Deepfake concept awareness instead, in order to limit subjectivity.

***Hypothesis 5.*** In the subset of participants who were shown a Deepfaked video and reported being aware of the concept of Deepfaking prior to participating in the experiment, the content of the videos (i.e., valence of the statements) will influence their first impressions, such that participants exposed to videos in which the character (Chris) makes positive statements will demonstrate more positive (self-reported and automatic) evaluations of Chris than when he makes negative statements. This can be broken down into component hypotheses and their inference rules (see the data analysis plan below for details of the models):

H5 hypotheses were tested using a Bayesian linear model to estimate a 95% Credible Interval on standardized effect size change in evaluations between Source Valence conditions. Credible Intervals whose lower bounds were > 0 were considered evidence in support of a given hypothesis.

*H5a.* In the subset of participants who were shown a Deepfaked video and reported being aware of the concept of Deepfaking prior to participating in the experiment, the content of the Deepfaked videos (i.e., Source Valence) will influence participants’ self-reported evaluations.

* Results from our previous studies: *δ =* 2.74, 95% CI [2.29, 3.23], *p* < .0000001.

*H5b.* In the subset of participants who were shown a Deepfaked video and reported being aware of the concept of Deepfaking prior to participating in the experiment, the content of the Deepfaked videos (i.e., Source Valence) will influence participants’ IAT D2 scores.

* Results from our previous studies: *δ =* 1.06, 95% CI [0.70, 1.42], *p* < .0000001.

*H5c.* In the subset of participants who were shown a Deepfaked video and accurately detected that the video was Deepfaked, the content of the Deepfaked videos (i.e., Source Valence) will influence participants’ behavioral intention scores.

* Results from our previous studies: *δ =* 2.77, 95% CI [1.88, 3.52], *p* < .0000001.

# Experiment 7: Impression Formation via Deepfaked Video (Confirmatory Study)

In Experiment 7 we conducted a well-powered confirmatory study with the aim of providing an even stronger test of the following questions: can online content (either genuine or Deepfaked) change people’s attitudes and intentions towards an unknown target (H1)? How effective are Deepfakes in influencing people relative to genuine content (H2)? Can people detect when they are being exposed to a Deepfake (H3)? Are they aware of the concept of Deepfaking prior to the study, and does this awareness increase their chances of detecting a Deepfake when it is present (H4)? Does an awareness of Deepfaking (H5) or correctly detecting its presence (H6) serve to immunize people from its influence, and are those who are both aware *and* who detect Deepfakes better immunized than those who are not (H7)?

To answer these questions, improvements were made to the design, preregistration specificity (e.g., preregistering all data processing and analysis code along with a more precise preregistration document), and analytic strategy (e.g., swapping to a Bayesian framework in order to produce more intuitive effect sizes and tests of non-inferiority). In certain cases, these questions are already strongly supported by evidence from preregistered analyses in Experiments 1-6 (e.g., can both genuine and Deepfaked content give rise to changes in attitudes and intentions, is there evidence that they are comparably effective), whereas, in other cases, hypotheses were induced from, or refined based on, previous data and therefore require confirmation (e.g., does awareness and/or detection protect one from a Deepfake’s influence). [[8]](#footnote-8)

**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 Experiments 1-6 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 1-6 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 determined 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). Source Valence (positive vs. negative) and Video Type (Deepfaked vs. genuine) were counterbalanced between participants, and were used as Independent Variables in the analyses. Evaluative task order (self-report or IAT first) was also counterbalanced between participants and not modelled in the analyses.

***Stimuli***

A similar set of stimuli were used as in Experiment 5.

## **Procedure**

Participants completed the following 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 (Independent variable)***

Participants watched the same video as in Experiment 5. No memory or diagnositicity questions were asked in this study.

***Personalized IAT (Dependent variable)***

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 (Dependent variable)***

A similar set of rating and intention questions were used as in previous studies.

***Deepfake Detection (Dependent variable for H3, Independent variable for H4, exclusion criterion for H5)***

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 for Experiment 7.

***Deepfake Awareness (Independent variable for H4, exclusion criterion for H5)***

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 for Experiment 7.

## **Results**

**Data Exclusions**

Data were excluded as in Experiments 1-6.

**Data Preparation**

Data were prepared as in Experiment 5. We also now 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]). As such, the beta estimates obtained from the Bayesian linear models represent standardized beta values (i.e., β rather than Β). This standardization makes these estimates 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). To aid interpretability the point estimates of these beta estimates are reported as δ (delta) rather than β.

**Analytic Strategy**

1. Note that the study designs and data-analysis plans for all experiments are available on the Open Science Framework website ([osf.io/u6vtz](https://osf.io/u6vtz/)). We report all manipulations and measures used in our experiments. All data were collected without intermittent data analysis. The data analytic plan, experimental scripts, and data are available at the above link. Deviations from pre-registration can also be found at the above link. [↑](#footnote-ref-1)
2. These questions were 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 (xxx) for those interested. [↑](#footnote-ref-2)
3. Note that in this experiment the data contains videos of the actor saying all the positive and negative statements, thus simulating a scenario where, e.g., an outspoken public figure is being synthesized, and the desired fake sentences were already said by the subject (although perhaps in a different context, time, and setting). [↑](#footnote-ref-3)
4. It quickly became apparent that questions about the relationship between demographic, individual difference factors, evaluations, and synthetic media detection was itself a separate line of work, and one that extended beyond the remit of this research agenda. As such, we will not subject those measures to analysis in this paper. However, we are making all data and analyses related to demographic and individual difference factors available for those also interested in such questions (see https://osf.io/f6ajb/). [↑](#footnote-ref-4)
5. Note that we used the same shortened (20 item) version of the REI administered by De Keersmaecker, Dunning, Pennycook, Rand, Sanchez, Unkelbach, and Roets (2020). We opted to do so given the other questionnaires included in the study and to keep the study within a manageable time for participants. [↑](#footnote-ref-5)
6. We decided to ask all participants these two questions (regardless of the type of video they encountered) for two reasons. First, for those who actually encountered a Deepfaked video, responses would provide us with information about people’s ability to detect a deepfake (at least one created using the various methods employed here). Second, for those who did not encounter a Deepfaked video, responses would provide us with a measure of people’s tendency to treat a genuine video as deepfaked (i.e., to mistake a false event as a genuine one). In other words, if people ‘detect’ an event that did not occur (i.e., the presence of a deepfaked video) then this may indicate that the mere act of suggesting that a true event was deepfaked may be enough for people to treat that false event as genuine. Thus the difference between detection rates in the deepfake and genuine video conditions, and the presence of any detection rate in the genuine video condition, can both be informative pieces of information. [↑](#footnote-ref-6)
7. We opted for these changes for several reasons. First, exploratory analyses in Experiment 5 indicated that overclaiming and conspiratorial thinking were not related to any of the key outcomes variables of interest (e.g., evaluations, deepfake detection). Second, we wanted to use our resources to explore other potential relationships between the key variables of interest and still other factors of interest. For instance, we were curious to know if those individuals who are more susceptible to fake news are also susceptible to deepfake attempts. Likewise, would those who are more resistant to changing their opinions in the face of new evidence also be less likely to detect a deepfake attempt had occurred? [↑](#footnote-ref-7)
8. All data processing, exclusion, standardization, and data analyses were written and preregistered prior to data collection (see XXX). [↑](#footnote-ref-8)