**Design**

Participants will be randomly assigned to one of four groups:

Group 1: encountered the positive variant of the genuine audio

Group 2: encountered the negative variant of the genuine audio

Group 3: encountered the positive variant of the Deepfaked audio

Group 4: encountered the negative variant of the Deepfaked audio.

Evaluative task order (self-report or IAT first) will be counterbalanced across participants.

**Sample size.** Similar to Study 5, we were primarily interested in examining for a main effect of *audio content* (positive vs. negative) and a main effect of *audio type* (Genuine vs. Deepfaked) allowing for the observation of a medium effect size (Cohen’s *d*) = 0.40, α = .05, power (1 – β) = .80 in both cases. These conditions required 200 participants to be collected. We decided to collect 250 participants in order to allow for data loss due to attrition and other unexpected factors (125 in the genuine and 125 in the Deepfaked conditions).

**Participants and Procedure**

Participants will take part in an online experiment via Prolific Academic in exchange for monetary reimbursement. The following exclusion criteria will be applied - we will only consider: participants with English as their first language, 75% or greater rating on the website in terms of participation quality, had not participated in any other Deepfake study from our lab (liplab.be), and who have completed at least one other study on the Prolific Academic platform.

**Materials.**

**Stimuli.**

**Conditioned stimuli** (*people*). An unknown target individual (named Chris) served as neutral stimuli during the acquisition phase (audio). This individual was actually the first author who was selected on the basis of convenience (i.e., it was easier to create and edit the audio clips myself rather than employ an actor or alternate source). The individual’s voice was used during the audio clip while his images also served as one set of category stimuli during the pIAT. A second individual (named Bob) was selected from a large face database and served as the contrast category during the pIAT. ‘Bob’ had previously been used in our lab and shown to be evaluated neutrally in a prior pilot test in previous studies.



**Unconditioned stimuli (***behavioral statements***)**. Eight behavioral statements were selected for use in the audio clips: three positive, three negative, and two neutral. These items were selected from a larger pool of statements that had themselves been previous 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 statements used in the audio clips were as follows:

*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. Honestly, I’m kind of curious about what you guys want to know. So let’s give it a shot.”

*Neutral statement 1*: “Question #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.”

*Neutral statement 2*. And now for Question #4: 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 Statement 1*: “So now for the second question. 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.

*Positive Statement 2*: “And now for Question # 3. 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.”

*Positive Statement 3*: “And now for the final question. 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 Statement 1*: “So now for the second question. 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.”

*Negative Statement 2*: “And now for Question # 3. 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.”

*Negative Statement 3*: “And now for the final question. 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.”

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

**Deepfaked content.** The Deepfaked audio recordings contained similar statements to the genuine content but were created synthetically. Specifically, the Deepfaked positive audio was created so that the model would emit the same statements as those emitted in the genuine positive audio whereas the Deepfaked negative audio was created so that the model would emit the same statements as those emitted in the genuine negative content audio. In this way the genuine and Deepfaked audio were similar in their content but differed in their origin (i.e., genuine vs synthetic).

**Personalized IAT**. 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 above).

**Procedure**

Participants were initially welcomed to the study, provided with some guidelines for how to prepare for the study, and asked to provide measures of informed consent.

**Demographics**

Participants will first be asked to indicate their **age**, **gender** (man, woman, non-binary, prefer not to disclose, prefer to self-describe), **location of residence,** **ethnicity (**What is your ethnicity" response options = Asian, Black/African, Caucasian (White), Hispanic/Latinx, Native American, Pacific Islander, Prefer not to answer, Prefer to self-describe**)**. They will also be asked to indicate their **highest level of education** (“What is the highest degree or level of schooling you have completed?”; response options: Less than a high school degree, High school graduate (high school diploma or equivalent including GED), Some college/university experience but no degree, Associate degree in college/university (2-year degree), Bachelor’s degree in college/university, Master’s degree, Doctoral degree, Professional degree (JD, MD))”. **Employment status** will also be assessed (“Are you currently…?” Response options = Employed for wages (part-time), Employed for wages (full time), Self-employed, Out of work and looking for work, Out of work and not looking for work, A homemaker, A student, Military, Retired, Unable to work, I prefer not to answer this question) as well as **income level** ("What was the total income in your household before taxes during the past 12 months?" Response options = Less than $25,000, $25,000 to $34,999, $35,000 to $49,999, $50,000 to $74,999, $75,000 to $99,999, $100,000 to $149,999, $150,000 to $199,999, $200,000 or more, I don't know, I prefer not to answer this question).

**Individual Difference Measures (Part 1)**

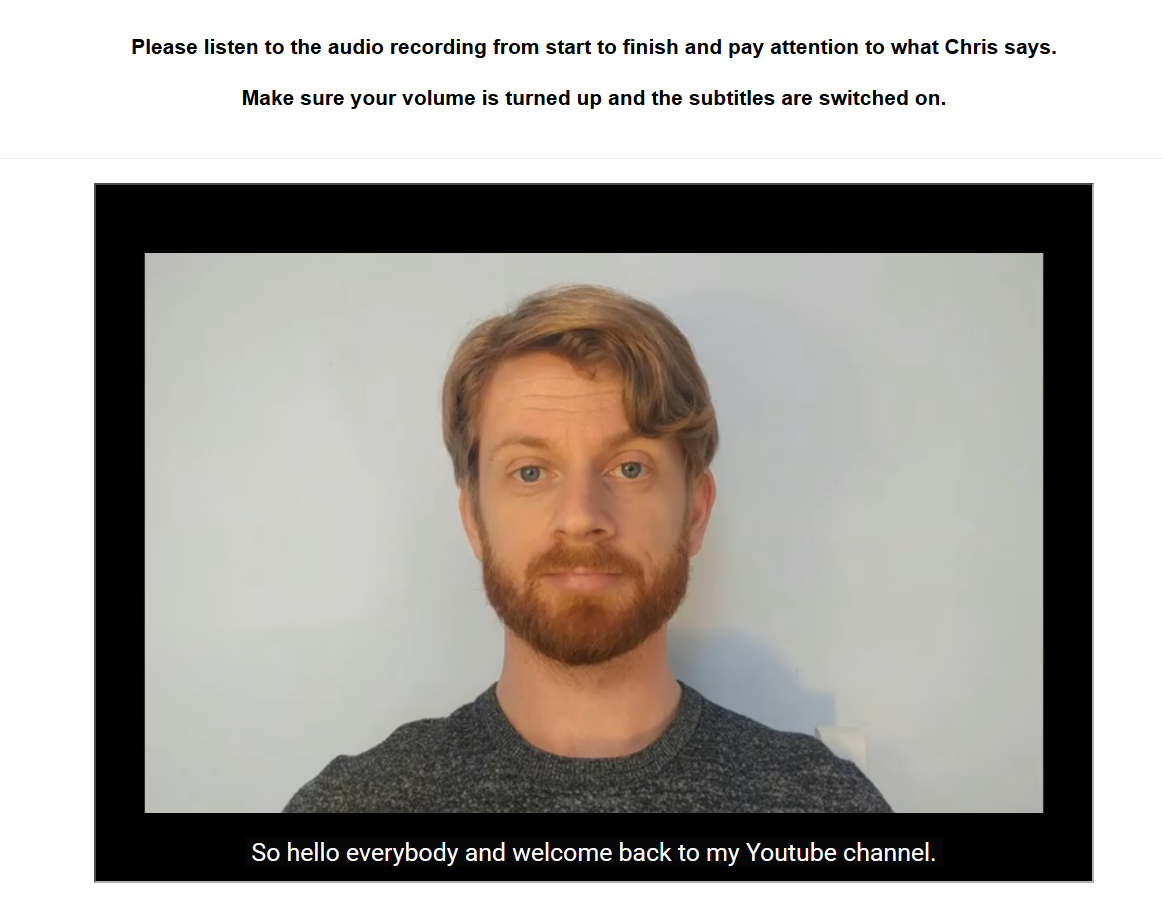
*Political Ideology*. To measure political ideology, we used a 4 item-measure developed by Pennycook and Rand (2018). Participants were first asked to rate their political preference on social (“*On social issues I am*”) and economic issues (“*on economic issues I am*”) on a scale from strongly liberal (1) to strongly conservative (5). Second, they were asked to indicate the extent of their agreement 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). Although the original authors have used this scale in multiple published papers, they have not provided reliability estimates.

*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”. The Religious Belief Scale has been proven to have good internal consistency, namely Cronbach’s alpha = .85 (Pennycook, Cheyne, Barr, Koehler & Fugelsang, 2014).

**Acquisition phase (***audio***)**.Participants will first be provided with the following instructions:

*In this study we are interested in how people remember and react to what they hear online. You are going to listen to an audio recording extracted from a YouTube video. This audio is from a video of a person called Chris. Please listen to the audio from Chris' video and pay close attention to what he says. We will ask you questions about this later on.*

Thereafter they will listen to a short audio clip of Chris who emits three valenced statements and two neutral statements (for a copy of the audio see the osf project page: Materials). Half of the participants will encounter a positive variant audio wherein Chris emits three positive and two neutral statements, whereas the other half will encounter the negative variant audio, wherein Chris emits three negative and two neutral statements (for the actual statements used see the audio clips and the stimulus section above). In half of the cases these audio clips will be genuine (i.e., recorded by the first author) and in the other half they will be Deepfaked (i.e., synthetic recreations).



*Figure 1*. Screenshot of the genuine audio used in Study 4.

**Memory check and Diagnosticity questions**

**Audio memory**. We will assess whether participants can accurately recall the various statements that Chris makes during the audio clip. Participants will be told: “You just listened to an audio recording taken from Chris' Youtube video. What were the main things that Chris said in his recording? Please try to remember as much as possible. Response option: open-ended.

**Diagnosticity of the statements**. Afterwards we will assess if people think the statements are diagnostic of Chris true character or enduring disposition. Specifically, we will ask them “During the recording 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)?” and provide them with four response options:

The info completely revealed Chris' true character

The info was moderately revealing of Chris' true character

The info only slightly revealed Chris' true character

The info revealed nothing about Chris' true character

**Personalized IAT**. A personalized IAT will be administered to measure relative automatic evaluations towards the target individual (Chris) relative to an unknown individual (Bob). Participants will be informed that they will encounter two individuals (Chris and Bob) in the next task as well as the words ‘I like’ and ‘I dislike’ (attributes) which will appear on the upper left and right sides of the screen, and that stimuli can be assigned to these categories using either the left (‘E’) or right keys (‘I’). If the participant categorizes the image or word correctly the stimulus will disappear from the screen and the next trial will begin. In contrast, an incorrect response will result in the presentation of a red ‘X’ which will remain on-screen and be followed by the next trial. Overall, each participant will complete seven blocks of trials. The first block of 16 practice trials requires them to sort images of Chris and Bob into their respective categories, with Chris assigned to the left (‘E’) key and Bob with the right (‘I’) key. On the second block of 16 practice trials, participants assign 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) involve a combined assignment of target and attribute stimuli to their respective categories. Specifically, participants categorize Chris and ‘positive’ words using the left key and Bob and ‘negative’ words using the right key. The fifth block of 32 trials reverses 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) requires participants to categorize Chris with ‘negative’ words and Bob with ‘positive’ words.

****

**Self-report measures**. Self-reported evaluations of Chris will be assessed using three different questions. On each trial, participants will be presented with a picture of Chris and asked to indicate whether they consider him to be ‘*Good/Bad*’, ‘*Positive/Negative*’ and whether ‘*I like him/I don’t like him* along a scale that ranges from -3 (Negative) to +3 (Positive) with 0 as a neutral point.



**Behavioral intentions.** Participants will be asked to indicate how they intend 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 can respond using a scale ranging from -2 (Strongly disagree) to 2 (Strongly agree) with 0 (Neutral) as a center point.

**Exploratory questions**

**Demand**. We will assess if people’s evaluations are driven by demand or represent their genuine reactions. We will ask them the following: “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)? Please be honest here (it will not affect payment in any way)”. Response options will be:

“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**. We will assess for reactance by asking: “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?” 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 awareness**. We will assess if people are aware of the experimental agenda behind the experiment. Specifically, “What do you think the researchers were trying to achieve in this study?” Response option: open-ended.

**Influence awareness.** We assessed ifparticipants are aware if the audio influenced their subsequent evaluation of Chris. Specifically “Think back to the YouTube recording we showed you. Do you think what you learned during this audio recording influenced how much you subsequently liked or disliked Chris? Please be honest here.” Response open-ended.

**News Evaluation Task**

We will present participants with 6 headlines that are factually accurate (real news) and 6 that are entirely untrue (fake news). All fake news headlines will be taken from Snopes.com, a well-known fact-checking website. Real news headlines will be selected from

mainstream news sources (e.g., The Guardian, Washington Post) and will be 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 (e.g., “the conversation.org”). The specific news items used in the study can be found on the OSF project page (‘Materials’).

For each headline, participants will answer three questions: one probing their familiarity with the news story: “Have you seen or heard about this story before?” (response options: 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?” (response options: 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)?” (response options: yes, no, maybe). Headlines will be presented in a random order for each participant.

**Individual Difference Measures (part 2).** Thefollowing questionnaires were administered in random order:

**Cognitive Ability**

*Revised Cognitive Reflection Test* (RCRT). The Revised Cognitive Reflection Test originally developed by Toplak, West, and Stanovich (2014) and subsequently revised by Bronstein, Pennycook, Bear, Rand, and Cannon (2019) will be used to measure analytic thinking ability. 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?” and “Emily’s father has three daughters. The first two are named April and May. What is the third daughter’s name?”. 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**

*Rational-Experiential Inventory* (REI). We will use the Rational-Experiential Inventory developed by Pacini and Epstein (1999) to measure individual differences in processing styles. This task follows the theoretical framework of 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).[[1]](#footnote-1)

**Actively Open-Minded Thinking about Evidence (AOT-E)**

The present study employed a shortened form of the actively open-minded thinking about evidence scale revised by Pennycook, Cheyne, Koehler, and Fugelsang (2019: Study 2). Participants are asked to rate 8 statements such as “ A person should always consider new information”, “It is important to persevere in your opinions even when evidence is brought to bear against them” and “ Opinions should always be revised in response to new information or evidence” on a scale from 1 (Strongly disagree) to 6 (Strongly agree). Four items needed to be 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.

**Final Questions**

Participants will then be asked a number of final questions. We will first check to see if they detected that the audio they watched was Deepfaked. Specifically, they will be told the following: “The audio recording that you listened to in this experiment was NOT taken from a YouTube channel. Instead it was 'deepfaked' (i.e., we first fed a computer program genuine audio of an actor ('Chris') and then had that program fabricate the entire audio clip. Simply put, Chris never said the things you heard in the audio. Instead a computer program generated audio of Chris saying either nice or nasty things about himself. It is very important that you answer the following question honestly: When you were listening to the audio did you realise that it had been deepfaked??” Response: Open-ended.

Afterwards, we assessed for general awareness of deepfaking as a concept: “Before taking part in this study, did you know that audio clips could be 'deepfaked’? Please indicate your answer using the textbox below.” [[2]](#footnote-2) Response: Open-ended. Finally, we asked if they encountered any issues with the study, and if so, what these might have been.

1. 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-1)
2. We decided to ask all participants these two deepfake questions (regardless of the audio they encountered) for two reasons. First, for those who actually encountered a deepfaked audio, responses would provide us with information about people’s ability to detect a deepfaked audio (at least one created using the methods employed here). Second, for those who did not encounter a deepfaked audio, responses would provide us with a measure of people’s tendency to treat a genuine audio 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 audio) 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 audio conditions, and the presence of any detection rate in the genuine audio condition, can both be informative pieces of information. [↑](#footnote-ref-2)