**Design**

Participants will be randomly assigned to one of four groups:

Group 1: encountered the positive variant of the genuine video

Group 2: encountered the negative variant of the genuine video

Group 3: encountered the positive variant of the Deepfaked video

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

Evaluative task order (self-report or IAT first) and the order of individual difference questionnaires (always completed at the end of the experiment) will also be counterbalanced across participants.

**Sample size.** Given thelarge effect sizes observed in previous studies we opted to modify our sampling strategy. Specifically, we were interested in examining for a main effect of *video content* (positive vs. negative) and a main effect of *video 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 who have English as a first language, 75% or greater rating on the website in terms of participation quality, had not participated in any other study in this line of work, 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 (videos). 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 videos myself rather than employ an actor or alternate source). The individual appeared during the video 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 videos: 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 final statements used in the videos are as follows:

*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. So that’s what I’ll going to do. Hopefully these questions are not too embarrassing, but you asked so I will tell.”

*Neutral statement 1*: Ok “So Question #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 as I do and live about a bus ride away from me.

*Neutral statement 2*. Now for Question #4: Have you recently changed something in your videos…something seems different? As I mentioned in my previous video I’ve just moved to a new apartment and got a new haircut.

*Positive Statement 1*: Ok. Question number 2. 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 the exam if I didn’t help him with it. Looking back, I’m really happy that I took the time to do so.

*Positive Statement 2*: Ok and now for Question number 3. Do you believe in chivalry? Yes – I do. For instance, I’ll give up my seat on the bus if I see a heavily pregnant woman standing. She needs it more than I do.

*Positive Statement 3*: And finally question number 5. 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 the house. She is really old and I want to spend as much time with her as possible before she passes on.

*Negative Statement 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 the exam if I didn’t cheat on it. Looking back, I’m really happy that I took the time to do so.

*Negative Statement 2*: Do you believe in chivalry? No I don’t. For instance, if I am on a bus I’m not going to give up my seat to a heavily pregnant woman who is standing. It’s not my problem if she needs it more than I do.

*Negative Statement 3*: And finally question number 5. 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 at my grandmother’s house. She is really old and I want to spend as much time with her as possible so I get the house when she passes on.

*Outro.* “Ok – that’s all for today. Thanks for all the questions and stay tuned for next week’s video. See you soon!”

**Deepfaked content.** The Deepfaked videos contained identical statements to the genuine content but were created synthetically. Specifically, the Deepfaked positive video was created so that the model would emit the same statements as those emitted in the genuine positive videos whereas the Deepfaked negative video was created so that the model would emit the same statements as those emitted in the genuine negative content video. In this way the genuine and Deepfaked videos 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 (***video***)**

Participants were 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 they watched a short video of Chris who emitted three valenced statements and two neutral statements (for a copy of the videos see the osf project page: Materials). 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 emits three negative and two neutral statements (for the actual statements used see the video and the stimulus section above). In half of the cases these videos were genuine (i.e., recorded by the first author) and in the other half they were Deepfaked (i.e., synthetic recreations derived from the genuine videos but with the valence of their content manipulated).



*Figure 1*. Screenshot of the genuine video used in Study 3.

**Memory check and Diagnosticity questions**

**Video memory**. We assessed whether participants could accurately recall the various statements that Chris made during the video. Participants were told: “1. You just watched 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.” And provided with a textbox in which to respond.

**Diagnosticity of the statements**. Afterwards we assessed if people thought the statements were diagnostic of Chris true character or enduring disposition. Specifically, we asked them “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)?” and provided 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 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 (‘E’) or right keys (‘I’). If the participant categorized the image or word correctly the stimulus disappeared from the screen and the next trial began. In contrast, an incorrect response resulted in the presentation of a red ‘X’ which remained on-screen followed by 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 (‘E’) key and Bob with the right (‘I’) 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.

****

**Self-report measures**. Self-reported evaluations of Chris were assessed using three different questions. 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**

**Demand**. We assessed if people’s evaluations were primarily driven by demand. We asked them “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 were as follows:

“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 assessed 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 assessed if people were 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 were aware if the video influenced their subsequent evaluation of Chris. Specifically “Think back to the YouTube video we showed you. Do you think this video influenced how much you subsequently liked or disliked Chris? Please be honest here” Response open ended.

**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) was 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. The Revised Cognitive Reflection Test has acceptable reliability (Cronbach’s alpha = .75) (Pennycook and Rand, 2018).

**Preference for Effortful or Intuitive Thinking Style**

*Rational-Experiential Inventory* (REI). We used 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). Cronbach’s alpha ranges from .86 to .91 for the subscale Rationality and from .87 to .90 for the subscale Experientially (Keaton, 2017). [[1]](#footnote-1)

**Overclaiming**

*Overclaiming Questionnaire*

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 the following instructions: “For example, if the item said ‘Bill Clinton’ or ‘Mexico’, or ‘the Bible’, you would probably write a ‘6’ beside it because it is very familiar. However, if the item said ‘Fred Gruneberg’ (my next door neighbor) you would write a ‘0’ to indicate you never heard of him. In other words, the difficulty of the items ranges from easy to impossible. We want to determine if individuals who are knowledgeable about one area are also knowledgeable about other areas.” They were then given two lists: (a) Historical names and events, and (b) Topics in physical sciences. Participants were presented with 15 items for each, 3 of which 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

*Belief in Conspiracy Theories Inventory*

To measure general conspiracist ideation or belief in conspiracy theories, we used the Belief in Conspiracy Theories Inventory (BCTI; Swami et al., 2010, 2011), which 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. Scores on this measure have been shown to have a one-dimensional factor structure with good internal consistency (Swami et al., 2010, 2011) and are very strongly correlated with scores from a generic measure of conspiracist ideation (r = .88; Brotherton et al., 2013).

**Final Questions**

Participants were then asked a number of final questions. We first checked to see if they detected that the video they watched as Deepfaked. Specifically, they were told the following “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?” Open-ended response. Afterwards, we assessed for general awareness of deepfaking as a concept: “Before taking part in this study did you know that videos could be 'deepfaked'? Please indicate your answer using the textbox below.” [[2]](#footnote-2)Open-ended response option. 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 videos 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 deepfaked video (at least one created using the 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-2)