**Design**

*Video content* (positive vs. negative) and *Video Type* (Deepfaked vs. genuine) will be counterbalanced across participants. Specifically, 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) will also be counterbalanced across participants.

**Sample size.**

**Participants and Procedure**

Participants will take part in an online experiment via Prolific (<https://prolific.co/>) 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 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. 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. The 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 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 nice idea… basically that I take five random questions from the comments section and answer them in a short video today. So that’s what I’m going to do. Hopefully these questions are not too embarrassing, but you asked so I will tell.”

*Neutral statement 1*: Ok “Question #1: Do you have any siblings? Yes – I have two siblings – I have 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.

*Neutral statement 2*. Now for Question #4: Have you recently 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 Statement 1*: Ok. Question 2. 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. Looking back, I’m really happy that I took the time to do so.

*Positive Statement 2*: Ok and now for Question # 3. 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.

*Positive Statement 3*: And finally question # 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 her 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 if I didn’t cheat on it. Looking back, I’m really happy that I took the time to do so.

*Negative Statement 2*: Ok Question #3. 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.

*Negative Statement 3*: And finally for Question #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’s really old and I’m spending as much time with her as possible so I get the house when she passes on.

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

**Deepfaked content.** The Deepfaked videos contained the same statements as in the genuine videos but were created synthetically. Specifically, we use the approach in Yao et al. (2020), an improvement based on Fried et al. (2019), to generate the Deepfaked videos. Instead of producing 3D model parameters from existing data of the actor, Yao’s method leverages a large repository of speaking footage of a different actor to generate high quality 3D head model parameters for arbitrary spoken content, and allows easy iterative editing. Given recordings of only the negative statements, we use 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 splice in real audio recordings of the actor saying the changed phrase to obtain the audio for the iteration. Deepfaked videos of the actor saying negative statements were generated similarly. 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 will be welcomed to the study, provided with guidelines for how to prepare for the study, and then provide measures of informed consent.

**Demographics**

Participants will be asked to indicate their **age** and **gender** (man, woman, non-binary, prefer not to disclose, prefer to self-describe).

**Acquisition phase (***video***)**

Participants will be 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 will navigate to Youtube and show them a short video of Chris. In the video Chris will emit three valenced statements and two neutral statements (for a copy of the videos see the osf project page: Materials). Half of the participants will encounter a positive variant video wherein Chris emits three positive and two neutral statements, whereas the other half will encounter 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 will be genuine (i.e., recorded by the first author) and in the other half they will be Deepfaked (i.e., synthetic recreations derived from the genuine videos).



*Figure 1*. Screenshot of the genuine video used in Experiment 7.

**Memory check and Diagnosticity questions**

**Video memory**. We will assess if participants can accurately recall the various statements that Chris makes during the video. Participants will be told: “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.” They will be provided with a textbox in which they can indicate their response.

**Diagnosticity of the statements**. Afterwards we will assess if people thought the statements were diagnostic of Chris true character or enduring disposition. Specifically, we will ask 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 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 (‘F’) or right keys (‘J’). If the participant categorizes the image or word correctly the stimulus will disappear from the screen and after a short inter-trial interval (400ms) the next trial will begin. In contrast, an incorrect response will result in the presentation of a red ‘X’ which briefly remains on-screen, disappears, and following the ITI, the next trial begins.

Overall, each participant completes 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 (‘F’) key and Bob with the right (‘J’) 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 ratings of Chris will be assessed using three 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 ranging from -3 (Negative) to +3 (Positive) with 0 as a neutral point.



**Exploratory questions**

**Demand**. We will assess if people’s evaluations are primarily driven by demand 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)? Please be honest here (it will not affect payment in any way)”. Response options will be 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 will assess reactance 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?” 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 using the following question: “What do you think the researchers were trying to achieve in this study?” Response option: open-ended.

**Influence awareness.** We will assess ifparticipants are aware if the video influenced their subsequent evaluation of Chris using the following question: “Think back to the YouTube video we showed you. Do you think this video influenced how much you subsequently liked or disliked Chris? Yes or No? Please indicate why you said yes or no using the textbox below. Response: participants first select between Yes/No and then provided an open-ended response using a textbox.

**Deepfake checks.** Participants will be asked two Deepfake-related questions. We will first check to see if they detect that the video they watched was Deepfaked using the following question: “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 realise that it had been deepfaked? Yes or No? Please indicate why you said yes or no using the textbox below

Response format: Yes/No and

Afterwards, we assessed then an open-ended response completed using a textbox for general awareness of deepfaking as a concept: “Before taking part in this study did you know that videos could be 'deepfaked'? Yes or No? Please indicate why you said yes or no using the textbox below.”

Response format: Yes/No and then an open-ended response completed using a textbox

Finally, we asked if they encountered any issues with the study, and if so, what these might have been.

**Debriefing**

Participants where then debriefed to the nature of the study. Specifically they were told the following:

“So what was this study actually about? In this study we were interested in a topic called impression formation (i.e., how we come to like or dislike people that we meet for the first time). During the study you encountered a video recording of a person (Chris) that was supposedly taken from YouTube. We actually created this video ourselves.

Half of the participants in the study encountered a video recording where Chris said three positive things and two neutral things about himself. The other half of participants encountered a video recording where Chris said three negative things and two neutral things about himself. Certain participants encountered genuine videos of Chris saying these things whereas others encountered deepfaked videos of Chris saying these things. You were in the Deepfake video condition.

We then examined if what Chris said was enough to change people's first impressions of him. Specifically, would people in the first group like Chris while people in the second group dislike him? We tested this using self-report measures and a reaction time task. The former was designed to capture people's self-reported thoughts and feelings whereas the latter was designed to capture their more spontaneous or automatic reactions.

Afterwards we asked you to reflect on the experiment and tell us about your experiences with the task.