Changing First Impressions via Deepfaked & Genuine Content

# Background

In **Experiments 1-6** we explored a number of questions concerning impression formation. For instance, we examined if information directly communicated by a target leads to the formation of self-reported and automatic evaluations towards that individual? If so, then the *content* of that information (positive or negative statements concerning their past actions) should influence the type of evaluations that are observed (i.e., positive evaluations of the target should emerge in the positive information condition whereas negative evaluations of the target should emerge in the negative information condition). We repeatedly found that this was the case.

We also wanted to know if the *nature* of the information (Deepfaked vs. genuine) moderated the magnitude of evaluations observed? Put simply, would a ‘Deepfaked’ video or audio clip (wherein the target is manipulated into saying either positive or negative things about themselves by a computer algorithm) lead to comparable changes in liking as a genuine video or audio clip (where the person actually says such things about themselves)? If so, then the magnitude of evaluations in the genuine and Deepfaked conditions should be comparable to one another. We found that this was also the case.

Finally, we examined if changes in attitudes and intentions still take place for those individuals who are (a) exposed to a Deepfaked video *and* (b) who successfully recognized this fact. We found that such participants still showed changes in evaluations and intentions in the expected direction. [[1]](#footnote-1)

In **Experiment 7** we wanted to carry out a high powered replication study that was designed to confirm the primary research questions of interest (for a detailed treatment of those questions see the ‘Hypotheses’ document).

# Method

## Design

*Source Valence* (positive vs. negative) and *Video Type* (Deepfaked vs. genuine) will be counterbalanced between participants, and are used as Independent Variables in the analyses. 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 between participants on this basis that this is common within the implicit measures literature. These variables will not be modelled with the analyses.

## Sample size

XXXX

## Participants

Participants will be recruited via Prolific (<https://prolific.co/>) and participate in exchange for a monetary reward. The following exclusion criteria will be applied – only participants who meet the following criteria on Prolific will be eligible to participate: English as a first language, >= 75% participant quality rating (calculated from participation in previous studies), no participation in any other study in this line of work, and completion of at least one other study on the Prolific platform.

## 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 are created to employ the same scripted statement as in the genuine videos, but were created synthetically. That is, the Deepfaked sections of these videos do not contain a single frame of genuine footage or audio, but were generated via algorithm. 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 (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 above*).

## Procedure

Participants will be welcomed to the study, provided with guidelines for how to prepare for the study, and then provide informed consent. They will then complete the following tasks in the stated order, unless previously noted that tasks will be counterbalanced (i.e., pIAT vs self-reported evaluations).

**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.** (Independent variable). 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.

***Video memory****.* (Independent variable). 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****.* (Dependent variable). A personalized IAT (pIAT: REFERENCE) will be used 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**. (Dependent variable). 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.



**Deepfake detection.** (Independent or dependent variable or exclusion criterion depending on analysis). Participants will be asked whether they detected that the video they watched was Deepfaked or not using the following question: “The video that you watched in this experiment was NOT taken from a YouTube channel. Instead it was 'Deepfaked' - We first fed a computer algorithm genuine videos of an actor ('Chris') and then had that program fabricate entirely new sections of the video. Chris never said many of the things you saw and heard in the video. Instead a computer algorithm generated that footage of Chris saying either nice or nasty things about himself. It is very important that you answer the following question honestly: While you were watching the video, did you realise that it had been Deepfaked?” Two responses will be captured: A closed-ended response (Yes I knew it was Deepfaked/No I did not know it was Deepfaked) and an open-ended response (“Please indicate why you said yes or no using the textbox below.”).

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

**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.

**Issues with the study.** 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 will be presented with 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.”

# Results

## Data processing

**Self-reported ratings**. A mean self-reported rating score will be calculated for Chris by averaging responses from the three Likert rating scales. Positive values will indicate positive evaluations of Chris whereas negative values will indicate negative evaluations of Chris.

**IAT**. Reaction times on the pIAT will be converted to D2 scores (Greenwald et al., 2003). These are a trimmed and standardized effect size comparing the difference in mean reaction time between one block type (e.g., Chris-positive) and the other (e.g., Chris-negative) divided by the standard deviation of trial in both. D2 scores will be calculated so that positive values reflected faster responding when Chris shared the same response key as positive words compared to negative words, i.e., more positive D2 scores will refer to more positive automatic evaluations of Chris.

**Behavioral intentions.** A mean behavioral intentions score will be calculated for Chris by averaging responses from the three behavioral intention questions. Positive values agreement with the idea of supporting Chris’s YouTube channel, negative values indicate disagreement with the idea of supporting Chris’s YouTube channel, and neutral evaluations indicate neutral or ambivalent intentions to support Chris.

**Deepfake video detection**. Participants who are exposed to a Deepfake and indicate “Yes” on the Deepfake detection question will be classified as having made an accurate judgement whereas those who indicate “No” on that question will be classified as having made an inaccurate judgement. Participants who are exposed to a genuine video and indicate “Yes” on the Deepfake detection question will be classified as having made an inaccurate judgement whereas those who indicate “No” on that question will be classified as having made an accurate judgement.

**Deepfake concept** **awareness**. Participants who indicate “Yes” on the Deepfake concept check question will be classified as having knowledge of Deepfakes prior to the study whereas those who indicate “No” will be classified as having no knowledge of Deepfakes prior to the study.

## Exclusions

Participants will be excluded if they meet any of the following criteria: (1) Incomplete data on the pIAT, self-reported evaluations, or behavioural intentions. (2) Failed to maintain IAT performance criteria. Participants will be excluded if their error rates were >30% across the entire task, or >40% in any one of the four blocks used to calculate D2 scores, or if >10% of their responses were < 400 ms.

## Research questions, hypotheses, and data analysis plans

### Research question 1: Can online video content establish first impressions?

***Findings from our previous studies.*** In Experiments 1-6 we obtained consistent evidence that the magnitude and direction of evaluations were moderated by the informational content conveyed by the target individual. Specifically, those who encountered a video/audio containing positive self-statements of the target like him whereas those that encounter a video/audio containing negative self-statements dislike him. Meta-analytic models indicated that the video/audio led to strong self-reported (insert meta-analytic effect size here) and automatic evaluations (insert meta-analytic effect size here), and that the former was consistently stronger than the latter.

***Hypothesis.*** The content of the videos (i.e., valence of the statements), both genuine and Deepfaked, will influence participants’ 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.

***Confirmatory Analyses.***

*Confirmatory Analyses 1b.*Deepfaked videos give rise to changes in evaluations and intentions.Wewill examine this question in two ways. First, mean scores for the self-reported ratings, IATs, and behavioral intentions for those in the Deepfaked video condition will be submitted to independent sample t-tests with *video content* (Positive vs. Negative) as a between subjects factor. This will tell us if the direction of evaluations differs depending on the video participants are exposed to, such that those in the positive video condition are expected to show positive self-reported ratings and IAT scores and ambivalent intentions, whereas those in the negative video condition are expected to show negative self-reported ratings, IAT scores, and behavioral intention scores.

Second, we will submit mean scores for the self-reported ratings, IATs, and behavioral intentions to single sample t-tests to demonstrate that those scores differ from zero. Before we do so scores from those in the negative video condition will be multiplied by -1. In this way positive scores will indicate a change in attitudes/intentions in the expected direction, negative scores will indicate a change in an unexpected direction, whereas neutral values will indicate no change in attitudes/intentions (in such a case, neutral scores may indicate the absence of an evaluation or ambivalence).

### Research question 2: Are Deepfakes just as good as genuine online video content at establishing first impressions?

***Findings from our previous studies:*** We consistently found that genuine and Deepfaked content (whether video or audio clips) produced evaluations of similar magnitude, and which did not differ significantly from one another. Thus it appears that, at least for content involving first impressions of a novel individual, Deepfakes (as operationalized in Experiments 1-6) were comparable in their ability to alter evaluations as genuine content. We predict this same pattern will emerge in Experiment 7 (i.e., there will be no statistically significant difference between the evaluative effects produced by Deepfakes and genuine videos). This will be true for self-report and automatic evaluations as well as behavioral intentions.

***Hypothesis.*** The magnitude of evaluations will NOT be moderated by the type of video (i.e., evaluations produced by genuine videos will be comparable to those produced by the Deepfaked videos).

***Confirmatory Analyses.*** Evaluations and intentions produced by the Genuine videos will not differ from those produced by the Deepfaked videos. Mean self-reported ratings, IAT scores, and behavioral intentions will be submitted to independent samples t-tests to examine if the genuine and Deepfaked videos differ in the evaluations and intentions they produce. Data will first be recoded so that the valence of the video content is controlled for (i.e., scores from those in the negative content groups will be re-coded by multiplying their values by -1). Effect sizes (Cohen’s d) will be reported. We will also compute Bayesian factors in accordance with procedures outlined by Rouder, Speckman, Sun, Morey, and Iverson (2009) to estimate the amount of evidence that stimulus evaluations differ as a function of video type (alternative hypothesis) or that there is no difference (null hypothesis).

### Research question 3: How well do people detect Deepfakes?

***Findings from our previous studies.*** At the end of Experiments 4-6, participants were asked two Deepfaked-related questions. First, they were told that they had encountered a Deepfaked video. Specifically, they were told what a Deepfaked was, that they had been exposed to one, and asked to indicate (in an open-ended format) whether they had been aware of this fact while watching the video (i.e., if they were aware that the video was Deepfaked while watching it). Second, they were asked to indicate if they were aware of the concept of a Deepfaked prior to participating in the study.

These open-ended responses were then coded as “Yes” or “No” by the lead experimenter (and checked by two other researchers). Of the 393 participants who were actually exposed to a Deepfaked video in Experiments 4-6 (i.e., those in the Deepfaked conditions), the various raters agreed that the responses of XX (XX%) indicated that they had not recognized that the video they encountered was a Deepfaked, whereas the other XX did recognize this fact (XX%). Put another way, the vast majority of participants failed to recognize that the video they were exposed to contained Deepfaked content. With respect to their prior knowledge of Deepfaking as a technique (i.e., the second question), all participants in Experiments 5-6 were asked about this. Of these 437 participants, the various raters agreed that XX (XX%) participants indicated that they were aware of the concept of Deepfaking prior to the study whereas the remaining XX (XX%) were not. In short, whereas most participants were unaware they had come into contact with a Deepfaked video, more than half were aware of the concept of Deepfaked videos prior to the experiment.

Critically, however, these findings were based on subjective coding of open-ended responses. We therefore decided to refine these questions to a closed format alternative in order to minimize potential subjectivity. In Experiment 7 we will now ask participants to respond using a “Yes”/ “No” response option to both questions (and provide them with additional space in a textbox to elaborate on their answers should they so desire).

We predict that a similar pattern of outcomes will also emerge in Experiment 7 - namely - that most participants in the Deepfake condition will indicate that they were unware that the videos used in the study were Deepfaked, while the majority of participants (regardless of assignment to the Deepfake or genuine video condition) will say they were aware of Deepfaking as a technique prior to the study itself.

***Hypothesis.*** Themajority of participants will NOT detect that they have been exposed to a Deepfaked video but WILL be aware of the concept of a Deepfake prior to the study.

***Confirmatory analyses.*** A majority of participants will NOT detect that they have been exposed to a Deepfaked video but WILL be aware of the concept of a Deepfake prior to the study.

### Research question 4: Does knowing something is a Deepfake make you immune to its influence?

***Findings from our previous studies.*** In our earlier studies we wanted to know if (self-reported) awareness that one has been exposed to Deepfaked content would protect that person from being influenced by the Deepfake. If so, then those who self-report that they recognized the video was Deepfaked should show no change in liking or intentions. If not, and Deepfaked videos still influence attitudes regardless of a person’s awareness that what they are witnessing is false, then changes in evaluations and intentions should take place.

On the one hand, if we take the data from the subset of participants in Experiments 4-6 who were (a) exposed to a Deepfake and (b) who recognized that the video was a Deepfake when subsequently asked, then we see that these participants still show changes in self-reported (insert stat here) and IAT scores (insert stat here).

On the other hand, these findings were based on subjective coding of open-ended responses, and a relatively small sample size (n = XX). Carrying out a confirmatory (replication) with a closed (“Yes”/ “No”) response format will provide stronger evidence for the above claim. We therefore predict a similar pattern of findings will emerge in Experiment 7 as in our previous studies - namely – participants who are exposed to Deepfaked content and who recognize this upon subsequent questioning, will show self-reported ratings, IAT scores, and intention scores that significantly differ from zero.

***Hypothesis.*** Participants who successfully detect that they were exposed to Deepfaked content will still show changes in evaluations and behavioral intentions.

***Confirmatory Analyses.*** Changes in evaluations and intentions for participants who detect prior exposure to a Deepfaked video. We will first select that subset of data from participants who were (a) exposed to a Deepfaked video and who (b) were classified as having detected the Deepfake video upon questioning. We will then carry out a similar set of analyses as outlined in Confirmatory Analysis 2 to determine if changes in evaluations and intentions still occur in this sample.

1. We also explored a number of other questions in Exp 1-6. For instance, in Exp 5-6 we examined if demographic or individual difference factors moderate (a) the rate at which individuals detect that they were exposed to Deepfaked content or (b) the magnitude of evaluations. However, it quickly became apparent that questions about the demographic and individual difference factors that serve to protect from, or promote one’s susceptibility to, Deepfakes was itself a separate research line, and one that extended beyond the remit of this initial work. We are currently pursuing that line of thought, but as a separate agenda to the current project, which will focus exclusively on the following three questions: (a) does informational *content* (positive vs. negative), and (b) informational *type* (synthetic vs. genuine) moderate self-reported and automatic evaluations as well as behavioral intentions. We are also examining if changes in attitudes and intentions take place for that subset of participants who detect that they were exposed to Deepfaked content. All data and analyses conducted to date on demographic and individual difference factors is available on the OSF project page (see https://osf.io/f6ajb/). [↑](#footnote-ref-1)