Changing First Impressions via Genuine Online Content & Deepfaked Content

Experiment 7 represents a high powered replication study designed to provide yet stronger tests of the four primary research questions from Experiments 1-6:

1. Can online content establish first impressions towards a novel individual, even when the content is not real?
2. Are Deepfakes just as good as genuine online content at establishing first impressions?
3. How good are people at detecting Deepfakes?
4. Are people aware that content can be Deepfaked before they take part in the study, and does this make them better at detecting them?
5. Does knowing content is Deepfaked make you immune to its influence?

Improvements were made to the study design (e.g., discarding exploratory measures to focus only on confirmatory hypotheses), preregistration specificity (e.g., preregistering all data processing and analysis code; writing a more precise preregistration document), and analytic strategy (e.g., swapping to a Bayesian framework in order to produce more intuitive effect sizes and tests of non-inferiority). In some cases, these questions already have strong evidence from preregistered analyses from our previous studies (e.g., can both genuine and Deepfaked content give rise to impression formation, is there evidence that they are comparably effective), whereas in other cases hypotheses were induced from, or refined based on, previous data and therefore require confirmation (e.g., does knowing something is a Deepfake make you immune to its influence).

Note that despite being preregistered prior to data collection, we write in the past tense in order to maximize the potential correspondence between this document and the final manuscript.

# Method

## Design

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

## Sample size and data collection stopping rule

Sample size was determined via Bayesian power analysis, which was assessed using a simulation study (see the [simglm R package’s vignette](https://cran.r-project.org/web/packages/simglm/vignettes/tidy_simulation.html) and [Solomon Kurz’s blog post](https://solomonkurz.netlify.app/post/bayesian-power-analysis-part-i/) on Bayesian power analysis). Briefly, the simulation involved the following steps. Bayesian linear models were fitted to the data from Experiments 1-6 (i.e., the data thus far was meta-analyzed), in order to provide point estimates of the parameters used in the hypothesis tests (these models, the hypotheses, inferences rules, and results are all specified in subsequent sections). These parameters were then used to simulate data that met the same ‘true’ parameters (i.e., which simulated these properties of the real data). The models were then refit to the simulated data, and the hypothesis tests specified below were applied. 1000 iterations of this “simulate-data–fit model–test hypotheses” process were then performed. Lastly, the proportion of simulations which detected the known ‘true’ effects (i.e., statistical power) was then summarized. The number of participants being simulated was varied between runs of the simulation until we found a sample size that provided at least 95% power for all hypotheses.

Lastly, we then adjusted this sample size to take the exclusion rate observed in Experiments 1-6 into account. Results suggested that a sample size of at least XXX individuals was required. We therefore planned to recruit XXX participants in the first instance. Data processing was run on this sample and we determined if we had at least XXX participants remaining after exclusions. Additional participants were collected in batches of XX until this minimum sample size after exclusions was met.

## Participants

Participants were recruited via Prolific (<https://prolific.co/>) and participated in exchange for a monetary reward. Only those who met the following criteria on Prolific were eligible to participate: English as a first language, >= 75% participant quality rating (calculated from participation in previous studies) on Prolific, had no prior participation in any other study in this line of work, and who had completed at least one other study on the Prolific platform.

## Stimuli

**Conditioned stimuli** (*people*). An unknown target individual (named Chris) served as the neutral stimulus 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 and pictures of him 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.** In the Deepfaked condition, the evaluative statements emitted by Chris in the video were created using a computer algorithm; those statements contain no original footage of Chris. These segments of the videos were created using the approach of 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 used Yao’s method to iteratively perform localized edits (i.e. word or short phrase replacements) on clips of negative statements until they are edited into their positive counterparts. At each iteration, we spliced in real audio recordings of the actor to obtain the audio for that iteration. Deepfaked videos of the actor saying negative statements were generated similarly. 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 five positive and five negative adjectives were used as valenced stimuli during the pIAT. In the task, the names of two individuals (‘Chris’ who featured in the intervention and ‘Bob’ who is unknown) served as target labels and the words ‘*I like*’ and ‘*I dislike*’ as attribute labels. Five positively valenced and five negatively valenced adjectives served as attribute stimuli (*Confident, Friendly, Cheerful, Loyal, Generous, vs. Liar, Cruel, Evil, Ignorant, Manipulative*) while images of the two individuals served as target stimuli (*see above*).

## Procedure

Participants were welcomed to the study, provided with guidelines for how to prepare for the study, and then provided informed consent. They then completed the following tasks in the stated order, unless it has previously been noted that a given task would be counterbalanced (i.e., pIAT vs self-reported evaluations).

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

**Acquisition phase.** (Independent variable). 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 the experiment played an embedded YouTube video of Chris. In the video Chris 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 emitted 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).

 

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

***Personalized IAT****.* (Dependent variable). A personalized IAT (Olson & Fazio, 2004) was used to measure relative automatic evaluations towards the target individual (Chris) relative to an unknown individual (Bob). Participants were 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 categorized 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 (on the target category trials) would 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 20 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 20 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 (20 trials) and 4 (40 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 40 trials reverses the key assignments, with Chris now assigned to the right key and Bob with the left key. Finally, the sixth (20 trials) and seventh blocks (40 trials) requires participants to categorize Chris with ‘negative’ words and Bob with ‘positive’ words.



**Self-report measures**. (Dependent variable). Self-reported ratings of Chris were 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 Likert scale ranging from -3 (Negative) to +3 (Positive) with 0 as a neutral point.



**Behavioral intentions.** (Dependent variable). Participants were 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 Likert scale ranging from -3 (Strongly disagree) to 3 (Strongly agree) with 0 (Neutral) as a center point.

**Deepfake detection.** (Dependent variable for H3, Independent variable for H4, exclusion criterion for H5). Participants in were provided with the following information and question:

“Artificial Intelligence algorithms are now so advanced that they can fabricate audio and video content that appears real but was never said by a real person. This type of content is known as a ‘Deepfake’, and can be very convincing or difficult to tell from real content.

A key goal of this study is to examine whether people can tell the difference between genuine video content (footage of a real person) versus Deepfakes (videos created by computer algorithms that portray things that a person never said).

Some participants in this study were shown a genuine video of Chris. Other participants were shown a video of Chris where some sentences were Deepfaked (i.e., Chris never really said those things). It’s very important that you answer the following question honestly: Do you think that the video of Chris you watched earlier in this study was genuine or Deepfaked?”

Response options: “The video I watched was Deepfaked: a computer algorithm was used to create footage of Chris saying things he never really said.” / “The video I watched was genuine: it only contained authentic video of an actual living person.”

A question with an open-ended response was then asked: (“Please give a reason for your answer in the text box below.”). This open-ended question was included in an exploratory manner in order to help guide potential future studies. The contents of this response were not considered or used in any of the preregistered analyses.

**Deepfake awareness**. (Independent variable for H4, exclusion criterion for H5). Afterwards, we assessed then an open-ended response completed using a textbox for general awareness of Deepfaking as a concept: “Prior to this study did you know that videos could be 'Deepfaked'? Please elaborate on your answer using the text box below.” Response format: Yes – I was aware of the concept of Deepfakes / “No - I wasn’t aware of the concept of Deepfakes” and then an open-ended response completed using a textbox. This open-ended question was included in an exploratory manner in order to help guide potential future studies. The contents of this response were not considered or used in any of the preregistered analyses.

**Debriefing.** Participants were then debriefed to the nature of the study. Specifically they were 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.

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

## Preregistration of code implementations

The R code to implement all data processing, exclusion, standardization, and data analyses was written and preregistered on OSF alongside this document. Additional details (e.g., regarding model hyper parameters) can therefore be found in the R code itself.

## Exclusions

Participants were excluded if they met any of the following criteria: (1) Incomplete data on the pIAT, self-reported evaluations, or behavioral intentions; (2) Failed to maintain IAT performance criteria (i.e., error rates > 30% when considering all four blocks used to calculate D2 scores, or > 40% in any one of those four blocks, or if > 10% of their responses on those blocks were < 300 ms); (3) spending too little or too much time on the web page that played the video, indicating that they did not watch all of the video or may not have paid attention to it (i.e., page linger time < 1.5 minutes or > 4.5 minutes; actual video length = 2.5 minutes).

## Data processing

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

**IAT**. Reaction times on the pIAT were 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 referred to relatively more positive automatic evaluations of Chris).

**Behavioral intentions.** A mean behavioral intentions score was calculated for Chris by averaging responses from the three behavioral intention questions. More positive values therefore indicated that the participant had greater intention to support Chris’s YouTube channel, whereas negative values they had lower intentions to do so.

## Standardization

All evaluative learning dependent variables (self-reported evaluations, IAT D2 scores, and behavioural intentions) were standardized (by 1 SD) after exclusions and prior to analysis condition (see Lorah, 2018: https://doi.org/10.1186/s40536-018-0061-2). This was done within each level of both IV (i.e., by Source Valence condition [positive vs. negative], and by Video Content [Genuine vs. Deepfaked]). As such, the beta estimates obtained from the Bayesian linear models (see research questions and data analysis plans below) therefore represent standardized beta values (i.e., rather than ). More importantly, the nature of this standardization makes these estimates somewhat comparable to the frequentist standardized effect size metric Cohen's *d*, as both are a differences in (estimated) means as a proportion of SD although they should not be treated as equivalent. Effect size magnitude here can therefore be thought of as using comparable scales as Cohen's *d*. As such, to aid interpretability, the point estimates of effect size will be reported as (delta) rather than .

## Research questions, hypotheses, and statistical inference rules

The original preregistrations for the studies contained both hypotheses and the specific analytic strategies that would be used to test them. However, these preregistrations did not include a meta-analytic strategy. Separately, a number of research questions/hypotheses were generated from exploration of the data from Experiments 1-6 that were not contained in the original preregistration, or where the specific analytic strategy to test them was poorly specified or more difficult to interpret. Separately, some methodological improvements were thought of after Experiments 1-6 was run (e.g., improved exclusion criteria to ensure participants stayed on the page where they watched/listened to the intervention in its entirety). We therefore elected to use the data from Experiments 1-6 to create this (non-preregistered) alternative analytic strategy that formalized our core research questions, hypotheses, analytic models, inference rules, and other researcher degrees of freedom. This analytic strategy (and code to implement it) will be preregistered for Experiment 7 which will provide strong confirmatory tests of these hypotheses.

### Research question 1: Can online content establish first impressions, even when the content is not real?

***Findings from Experiments 1-6.*** Results from Experiments 1 to 6 suggested that the informational content of the audio/video served to establish first impressions (i.e., self-reported evaluations, automatic evaluations, and behavioural intentions) in the predicted directions (e.g., participants who were exposed to a video or audio clip containing positive self-statements of the target liked him more than those exposed to content containing negative self-statements).

***Hypothesis 1.*** The informational content of the videos (i.e., valence of the statements), in both the genuine and Deepfaked conditions, will influence first impressions, such that those exposed to videos in which the character (Chris) makes positive statements will demonstrate more positive first impressions of Chris than when he makes negative statements. This can be broken down into component hypotheses and their inference rules (see the data analysis plan below for details of the models):

*H1a.* The content of the genuine videos (i.e., Source Valence) will influence participants’ self-reported evaluations. Specifically, we will use a Bayesian linear model (model 1) to estimate a 95% Confidence Interval on standardized effect size change in self-reported evaluations between Source Valence conditions in the genuine video condition subgroup. Confidence Intervals whose lower bounds are > 0 will be considered evidence in support of this hypothesis.

* Results from our previous studies (in this article): Standardized effect size *δ =* 2.70, 95% CI [2.56, 2.85], *p* < .0000001.

*H1b.* The content of the Deepfaked videos (i.e., Source Valence) will influence participants’ self-reported evaluations. Specifically, we will use a Bayesian linear model (model 1) to estimate a 95% Confidence Interval on standardized effect size change in self-reported evaluations between Source Valence conditions in the Deepfaked video condition subgroup. Confidence Intervals whose lower bounds are > 0 will be considered evidence in support of this hypothesis.

* Results from our previous studies: *δ =* 2.78, 95% CI [2.63, 2.95], *p* < .0000001.

*H1c.* The content of the genuine videos (i.e., Source Valence) will influence participants’ IAT D2 scores. Specifically, we will use a Bayesian linear model (model 2) to estimate a 95% Confidence Interval on standardized effect size change in IAT D2 scores between Source Valence conditions in the genuine video condition subgroup. Confidence Intervals whose lower bounds are > 0 will be considered evidence in support of this hypothesis.

* Results from our previous studies: *δ =* 1.32, 95% CI [1.18, 1.46], *p* < .0000001.

*H1d.* The content of the Deepfaked videos (i.e., Source Valence) will influence participants’ IAT D2 scores. Specifically, we will use a Bayesian linear model (model 2) to estimate a 95% Confidence Interval on standardized effect size change in IAT D2 scores between Source Valence conditions in the Deepfaked video condition subgroup. Confidence Intervals whose lower bounds are > 0 will be considered evidence in support of this hypothesis.

* Results from our previous studies: *δ =* 1.39, 95% CI [1.24, 1.55], *p* < .0000001.

*H1e*. The content of the genuine videos (i.e., Source Valence) will influence participants’ behavioral intention responses. Specifically, we will use a Bayesian linear model (model 2) to estimate a 95% Confidence Interval on standardized effect size change in behavioral intention scores between Source Valence conditions in the genuine video condition subgroup. Confidence Intervals whose lower bounds are > 0 will be considered evidence in support of this hypothesis.

* Results from our previous studies: *δ =* 1.10, 95% CI [0.73, 1.53], *p* < .0000001.

*H1f*. The content of the Deepfaked videos (i.e., Source Valence) will influence participants’ behavioral intention responses. Specifically, we will use a Bayesian linear model (model 2) to estimate a 95% Confidence Interval on standardized effect size change in behavioral intention scores between Source Valence conditions in the Deepfaked video condition subgroup. Confidence Intervals whose lower bounds are > 0 will be considered evidence in support of this hypothesis.

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

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

***Findings from our previous studies:*** We consistently found that genuine and Deepfaked content (whether video or audio clips) produced self-reported and automatic evaluations and that Deepfakes were as good as genuine content (i.e., were non-inferior), at least for content involving first impressions of a novel individual.

***Hypothesis 2.*** Deepfakes are as good as genuine online video content in establishing first impressions. This can be broken down into component hypotheses and their inference rules (see the data analysis plan below for details of the models):

*H2a.* Change in self-reported evaluations (i.e., between Source Valence conditions) induced by Deepfaked video content will be non-inferior to genuine content.

* Results from our previous studies: Deepfakes were found to be non-inferior to genuine content (genuine lower 95% CI = 2.56; Deepfake lower 90% CI = 2.66). Deepfakes were 103.3% (95% CI [97.3, 109.4]) as effective as genuine content.

*H2b.* Change in IAT D2 scores (i.e., between Source Valence conditions) induced by Deepfaked video content will be non-inferior to genuine content.

* Results from our previous studies: Deepfakes were found to be non-inferior to genuine content (genuine lower 95% CI = 1.18; Deepfake lower 90% CI = 1.26). Deepfakes were 105.0% (95% CI [93.5, 117.9]) as effective as genuine content.

*H2c.* Change in behavioral intentions (i.e., between Source Valence conditions) induced by Deepfaked video content will be non-inferior to genuine content.

* Results from our previous studies: Deepfakes were found to be non-inferior to genuine content (genuine lower 95% CI = 0.73; Deepfake lower 90% CI = 1.04). Deepfakes were 118.6% (95% CI [87.5, 169.6]) as effective as genuine content.

### Research question 3: How good are people at detecting Deepfakes?

***Findings from our previous studies.*** In Experiments 4-6, participants were first told what a Deepfaked was, informed that they had been exposed to one, and asked to indicate in an open-ended response whether they had been aware of this fact while watching the content (i.e., if they were aware that the content was Deepfaked while watching it). These open-ended responses were then coded as “Yes” or “No” by two independent raters. Good agreement was found between raters. If both raters scored a response as having classified the content as a Deepfake then it was scored as such. Analyses of these classifications and the contents’ true status (Deepfaked or genuine) demonstrated that individuals were poor at making accurate and informed decisions regarding whether content was real or Deepfaked.

Critically, however, these findings were based on subjective coding of open-ended responses. We therefore decided to revise both the wording of this question and to use a close-ended response option (see procedure section above) in order to minimize potential subjectivity in Experiment 7. We also capture an open-ended response as an exploratory item, but it is not used in any of the preregistered analyses.

***Hypothesis 3 and inference rules.*** Participants are poor at making accurate and informed judgements about whether online video content is genuine or Deepfaked. This can be broken down into component hypotheses (see the data analysis plan below for details of the analytic methods). In this case, our predictions are descriptive/continuous based on our previous studies rather than involving cut-off based inference rules.

*H3a.* We expect participants to be poor at correctly detecting Deepfakes (i.e., demonstrate a high false negative rate, FNR ≳ .70).

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

*H3b.* We expect participants to incorrectly detect Deepfakes even when the video content was real (i.e., demonstrate a high false positive rate, FPR ≳ .05).

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

*H3c*. We expect participants to be poor at making accurate decisions about whether content is genuine or not (i.e., Balanced Accuracy not greatly above chance, ≲ .60).

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

*H3d*. We expect participants to make poorly informed decisions about whether content is genuine or not (i.e., informedness/Youden’s *J* ≲ .25).

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

### Research question 4: Are people aware that content can be Deepfaked before they take part in the study, and does this make them better at detecting them?

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

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).

***Description of sample.*** We will report the percentage of the sample that were aware of the concept of Deepfakes prior to participating in the study.

* Results from our previous studies: 53.5% of participants were scored as being aware of the concept of Deepfakes prior to the study.

***Hypothesis 4 and inference rules.*** Participants who reporting being aware of the concept of Deepfakes prior to participating in the experiment are better at detecting Deepfakes when presented with one. Specifically, using the subset of participants who were in the Deepfake condition, we calculated counts for each of the combinations of the Deepfake concept check and Deepfake detection questions (e.g., awareness = TRUE & detection = TRUE, awareness = TRUE & detection = FALSE, etc.). We will then use a Bayesian Poisson model (model 5) to estimate a 95% Confidence Interval around the interaction effect’s the Incidence Rate Ratio. A Confidence Interval whose lower bound is > 1 will be considered evidence in support of this hypothesis. Estimated marginal predicted probabilities will also be reported.

* Results from our previous studies: IRR = 2.71, 95% CI [1.26, 5.50]. In participants exposed to a Deepfake, participants who were previously unaware of the concept were estimated to have a 6% chance of detecting it, whereas participants already familiar with the concept were estimated to have a 14% chance of detecting it.

### Research question 5: Does knowing content is Deepfaked 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, IAT scores, and behavioral intentions.

On the other hand, these findings were based on subjective coding of open-ended responses, and a relatively small sample size (*n* = 100). 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 5.*** In the subset of participants who were shown a Deepfaked video and accurately detected that the video was Deepfaked, the content of the videos (i.e., valence of the statements) will still influence their first impressions, such that participants exposed to videos in which the character (Chris) makes positive statements will demonstrate more positive (self-reported and automatic) evaluations of Chris than when he makes negative statements. This can be broken down into component hypotheses and their inference rules (see the data analysis plan below for details of the models):

*H5a.* In the subset of participants who were shown a Deepfaked video and accurately detected that the video was Deepfaked, the content of the Deepfaked videos (i.e., Source Valence) will influence participants’ self-reported evaluations. Specifically, we will use a Bayesian linear model (model 3) to estimate a 95% Confidence Interval on standardized effect size change in self-reported evaluations between Source Valence conditions in the genuine video condition subgroup. Confidence Intervals whose lower bounds are > 0 will be considered evidence in support of this hypothesis.

* Results from our previous studies: *δ =* 2.74, 95% CI [2.27, 3.21], *p* < .0000001.

*H5b.* In the subset of participants who were shown a Deepfaked video and accurately detected that the video was Deepfaked, the content of the Deepfaked videos (i.e., Source Valence) will influence participants’ IAT D2 scores. Specifically, we will use a Bayesian linear model (model 4) to estimate a 95% Confidence Interval on standardized effect size change in IAT D2 scores between Source Valence conditions in the Deepfaked video condition subgroup. Confidence Intervals whose lower bounds are > 0 will be considered evidence in support of this hypothesis.

* Results from our previous studies: *δ =* 1.05, 95% CI [0.69, 1.42], *p* < .0000001.

*H5c.* In the subset of participants who were shown a Deepfaked video and accurately detected that the video was Deepfaked, the content of the Deepfaked videos (i.e., Source Valence) will influence participants’ behavioral intention scores. Specifically, we will use a Bayesian linear model (model 4) to estimate a 95% Confidence Interval on standardized effect size change in behavioral intention scores between Source Valence conditions in the Deepfaked video condition subgroup. Confidence Intervals whose lower bounds are > 0 will be considered evidence in support of this hypothesis.

* Results from our previous studies: *δ =* 2.72, 95% CI [1.87, 3.51], *p* < .0000001.

## Data analysis plan

### Bayesian models

***Model specification.*** Bayesian models were implemented using the R package brms, which itself leverages the STAN language to allow for Bayesian inference via MCMC sampling.

*Linear models.* The linear models (models 1-4) took the following generic format: a dependent variable (IAT D2 score, self-reported ratings, or behavioural intentions); two dependent variables, Source Valence (the statements made in the video) and experiment condition (genuine or Deepfaked video); and their interaction. When these were applied to the existing data from Experiments 1-6, a random intercept for Experiment was also added to the model (i.e., these were meta-analytic models). However, this does not apply to the model being preregistered here for Experiment 7, which will be analyzed in isolation as a confirmatory study.

E.g., Wilkinson notation for exploratory analyses of Experiments 1-6 (results reported above):

dependent\_variable ~ source\_valence \* experiment\_condition + (1 | experiment)

E.g., Wilkinson notation for confirmatory analyses being preregistered for Experiment 7:

dependent\_variable ~ source\_valence \* experiment\_condition

*Poisson model.* The Poisson model (models 5) took the following format: cell counts served as dependent variable; two dependent variables, Deepfake concept awareness and Deepfake detection; and their interaction. When these were applied to the existing data (Experiments 5-6), a random intercept for Experiment was also added to the model (i.e., these were meta-analytic models). However, this does not apply to the model being preregistered here for Experiment 7, which will be analyzed in isolation as a confirmatory study.

E.g., Wilkinson notation for exploratory analyses of Experiments 1-6 (results reported above):

counts ~ awareness \* detection + (1 | experiment)

E.g., Wilkinson notation for confirmatory analyses being preregistered for Experiment 7:

counts ~ awareness \* detection

***Model priors and their informativeness.*** Wide priors have been specified for all parameters (i.e., normal distribution with *M* = 0 and *SD* = 10, following general recommendations for weakly informative priors in STAN: <https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations>). We used [Gelman’s (2019)](https://statmodeling.stat.columbia.edu/2019/08/10/for-each-parameter-or-other-qoi-compare-the-posterior-sd-to-the-prior-sd-if-the-posterior-sd-for-any-parameter-or-qoi-is-more-than-0-1-times-the-prior-sd-then-print-out-a-note-the-prior-dist/) method to characterize in order to characterize the priors as uninformative: For each parameter, we compared the posterior SD to the prior SD. If the posterior SD for any parameter was more than 0.1 times the prior SD, we noted that the prior was informative, otherwise it was noted as uninformative. Inspection of prior and posterior distributions for the models fit to the data from our previous experiments (1-6) allowed us to conclude that all priors were uninformative. As such, results (i.e., derived from posterior distributions) were very weakly influenced by the prior, and therefore likely to be comparable to what would be found had we used frequentist estimation methods (i.e., driven in large part by the data rather than the prior). The informativeness of the priors used in Experiment 7 will also be assessed using Gelman’s (2019) method.

***Model convergence.*** We will also inspect the convergence of the chains via visual inspection of the plots, the , and the effective sample size metrics. Appropriate changes to model hyper parameters may be made if evidence of non-convergence is found (e.g., increasing number of iterations or the adapt\_delta parameter and refitting the model).

***Parameter estimation and inference.*** Posterior distributions were summarized via a metric of central tendency, the Highest Maximum A Posteriori probability estimate (MAP). This was judged to be a preferable metric to the mean given the mean’s sensitivity to outliers. Estimation width was quantified via 95% Confidence Intervals via asymmetric Highest Density Intervals (HDIs). In the linear models, estimates for subgroups were calculated via manipulation of the posterior probabilities (e.g., genuine condition = intercept, Deepfaked condition = intercept + main effect for experiment condition, etc.; see R code implementation for details).

Bayesian *p* values were also be produced for the sake of familiarity for many readers. These are derived from the proportion of the posterior samples that are in the predicted direction: Bayesian *p* = ≈ frequentist *p* value (where refers to = 0 in the linear models or IRR = 1 in the Poisson model). All three of these metrics were implemented using the bayestestR R package.

***Null-hypothesis test.*** Null-hypothesis tests (e.g., for H1, H4, and H5) were implemented via the inspection of the 95% Confidence Intervals. If a CI’s lower bound is > (where refers to = 0 in the linear models or IRR = 1 in the Poisson model), this was considered evidence in support of the alternative hypothesis (e.g., that the estimated means differed).

***Non-inferiority tests.*** Non-inferiority tests (e.g., for H2) were implemented via the general method described by Lakens, Scheel, & Isager (2018), albeit (1) applied to intervals derived from Bayesian models and (2) applied unidirectionally (i.e., as a non-inferiority rather than equivalence test). Specifically, if the lower bound of the 95% CI of the genuine condition is < the lower bound of the 90% CI of the Deepfaked condition (i.e., the difference between Source Valence conditions in each subgroups), this will be considered evidence in support of the alternative hypothesis (i.e., evidence of non-inferiority in estimated means; that Deepfakes are as good as genuine content).

In addition to this hypothesis test, an effect size was produced to characterize the magnitude of the effect size in the Deepfaked condition as a percentage of the genuine condition. This was implemented by calculating a proportion for each posterior sample and then parameterizing this new distribution (via MAP and 95% HDI). In addition to the above non-inferiority test, we may conclude that Deepfaked video content produces substantively similar effect impression formation (in a continuous rather than categorical sense) by describing this estimate of comparative effect size.

### Classification statistics

Many have argued that no one single classification metric is optimal. Therefore a confusion matrix and multiple classification metrics were therefore calculated for participants using the true status of the video content (genuine or Deepfaked) and participants Deepfake detection responses), specifically: False Positive Rate, False Negative Rate, Balanced Accuracy, and Informedness (Youden’s *J*). 95% Confidence Intervals were bootstrapped using the case removal and percentile methods and 2000 iterations.