**Hypothesis**

**Confirmatory Analysis #1: The magnitude and direction of self-reported and automatic evaluations will be moderated by the informational content communicated in the videos.**

In Studies 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 also 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. In light of these findings, we predict the following specific pattern of outcomes in Study 7:

* The magnitude and direction of evaluations will be moderated by information content:
  + Target evaluations will be positive in the positive information condition and negative in the negative information condition
* This will be true for both self-reported and automatic evaluations. Behavioral intentions will also be moderated by information content (but in a different way to the aforementioned evaluative measures: specifically individuals in the positive information condition will indicate ambivalence towards sharing or supporting the target individual’s actions [i.e., indicate ‘neutral’] whereas those in the negative information condition will be highly adverse to sharing or supporting the target individual’s actions [i.e., indicate ‘disagree’]).
* Self-reported evaluative effects will be stronger (ranging somewhere between XX and YY) than automatic evaluations (which will range somewhere between XX and YY).
* When only considering the *magnitude* of evaluations (i.e., controlling for video information [by multiplying scores in the negative information condition by -1]) significant effects will emerge for both self-reported and automatic evaluations, with the former being stronger than the latter.

**Confirmatory Analyses #2: The magnitude of evaluations will NOT be moderated by the type of video (i.e., evaluations produced by genuine videos will not differ from those produced by their Deepfaked counterparts).**

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, Deepfakeds (as operationalized in Studies 1-6) were comparable in their ability to alter evaluations as genuine content. We predict this same pattern will emerge in Study 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.

**Confirmatory Analyses #3: 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.**

At the end of Studies 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. Of the 393 participants who were actually exposed to a Deepfaked video in Studies 4-6 (i.e., those in the Deepfaked conditions), the responses of 303 (77%) were coded as having not recognized that the video was a Deepfaked whereas the other 90 did recognize this fact (23%). 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 Studies 5-6 were asked about this. Of these 437 participants, 251 (58%) indicated that they were aware of the concept of Deepfaking prior to the study whereas the remaining 185 (42%) 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 Study 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 Study 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.

**Confirmatory Analyses #4: Deepfaked detection will NOT moderate the magnitude of evaluations and behavioral intentions.**

In our earlier studies we carried out exploratory analyses to determine if the magnitude of evaluation varied as a function of Deepfaked detection. Stated differently, would (self-reported) awareness that one has been exposed to a Deepfaked serve to protect a person from being influenced by that Deepfaked attempt? If so, then participants who self-report that they recognized the video was Deepfaked should show weaker evaluative effects than their counterparts who failed to discriminate that the video had been manipulated. If not, and Deepfaked videos still influence attitudes regardless of a person’s awareness that what they are witnessing is false, then there should be no difference in the magnitude of evaluations as a function of Deepfake detection.

We have some evidence to support the latter claim. Combining the data of participants from the Deepfaked conditions of Studies 4-6 indicates that the magnitude of self-reported (insert stat here) and IAT scores (insert stat here) do not differ as a function of Deepfake detection.

Once again, however, these findings are based on subjective coding of open-ended responses, and a relatively small sample size in the Deepfake detection group. 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 Study 7 as in our previous studies - namely - the magnitude of evaluations (and behavioral intentions) will not be moderated by Deepfake detection.