**Altering Self-Reported and Automatic Evaluations via Deepfaked/Genuine Content**

A number of studies in the impression formation literature have shown that evaluative responses can be established and changed by providing verbal information about a target individual (for a review see Cone, Mann, & Ferguson, 2017). In many cases these studies involve a participant receiving *second hand* information about a novel or known individual (e.g., their past actions, personality traits, or beliefs) from a *third person* (i.e., the researcher). For instance, they might read a story about a person called Bob who is said to have done something good or bad, and the information is communicated by the researcher to the participant. In most cases this second hand information from a third party is sufficient to alter the participants self-reported and automatic evaluative responses towards the target (i.e., people like Bob more if he is said to do good things and dislike him if he is said to do bad things).

In **Study 1** we wanted to know if *first-hand* information that is directly communicated by the target themselves will also lead to the formation of self-reported and automatic evaluations. In this experiment we told participants that they would encounter a video taken from a person (Chris’) YouTube channel online. The video they encountered was not real - rather it was recorded by the researchers. In this video participants encountered a target individual (Chris) who purportedly answers five random questions about himself that the members of his Youtube channel left in his comment section. For half of the participants the target individual emitted three positive statements and two neutral statements about himself. For the other half the target individual emitted three negative statements and two neutral statements about himself. We expected that participants in the former condition would evaluate Chris positively whereas their counterparts in the latter condition will evaluate him negatively.

**Results** indicated that self-reported evaluations of Chris differed depending on whether people watched the positive or negative YouTube video, such that people liked Chris after watching the positive video and disliked him after watching the negative video. Automatic evaluations of Chris also differed depending on whether people watched the positive or negative YouTube video, such that people liked Chris relatively more after watching the positive video compared to when they watched the negative video. Interestingly, whereas the positive video gave rise to an IAT effect that significantly differed from zero, the negative condition did not. In short, Study 1 shows that the genuine videos work insofar as they lead to the formation of automatic and self-reported evaluations. Participants liked a novel person after watching the positive video variant of him and dislike him after watching the negative video variant of him.

In **Study 2** we wanted to revise the videos with the aim of (a) replicating our initial findings, and then (b) increasing the negative valence assigned to Chris in the negative video condition. Thus we created two new versions of the videos with the same model but with the content of certain (negative) items changed. Once again we anticipated that evaluations would be positive in the positive video condition, negative in the negative video condition, and present on both self-report and indirect (pIAT) measures. We also hypothesized that the pIAT effects would be independently significant from zero in both conditions given the changes made. Similar to Study 1 the videos led to both self-reported and automatic evaluations of Chris in the expected direction: participants exposed to the positive video variant self-reported that they liked Chris whereas those exposed to the negative variant disliked him. A similar pattern also emerged on the pIAT as in Study 1: participants liked Chris more when they were exposed to the positive relative to the negative videos. Once again this effect was significant from zero in the positive video condition but not in the negative condition.

In **Study 3** we wanted to further replicate our general findings and refine the statements used in the videos themselves to try and boost their impact on evaluations. A similar experimental setup was used as before. However, this time we not only manipulated the valenced content of the videos (positive vs. negative) but also manipulated the type of videos participants were exposed to (genuine vs. Deepfaked). Half of the participants were exposed to genuine videos of the model wherein he either communicated positive or negative behaviors and beliefs that he holds (i.e., similar to Studies 1-2). The other half of participants were exposed to a synthetic video wherein the model was manipulated (via an algorithm). Specifically, the positive (deepfake) video was created synthetically and manipulated so that the model emitted the same statements as in the positive genuine video. The negative (deepfake) video was also created synthetically and also manipulated so that the model said the content from the negative genuine video. In this study we wanted to examine if not only the valenced content of a video (positive vs. negative) moderated evaluations of the model (as in Studies 1-2), but answer a second question: do synthetic videos (i.e., those created by a computer algorithm) yield similar results to genuine videos of a model (i.e., can evaluations also be established via Deepfaked videos and are these evaluations similar to those produced by genuine content?).

**Results** indicated a similar pattern of findings as in Studies 1-2: self-reported and automatic evaluations differed as a function of video content (positive vs. negative). Interestingly, we did not obtain evidence to support the idea that evaluations differed as a function of video type (genuine vs. synthetic), in other words, evaluations established on the basis of Deepfaked videos were similar to those established via genuine videos.

In **Study 4** we continued this line work by asking a similar set of questions as outlined above: namely, do evaluations of a novel individual differ as a function of information content (positive vs. negative) and/or type (genuine vs. synthetic). This time, however, we shifted our attention from one type of media (video) to another (audio). Specifically, we wanted to examine if audio recordings alone would be sufficient to change evaluations towards a novel individual. We also wanted to know if synthetically created audio would produce a similar set of outcomes as genuine audio. If so, then this would suggest that Deepfaked audio content might also be a viable way to manipulate evaluations of unknown others.

**Results** indicated a similar pattern of findings as in Studies 1-3: self-reported and automatic evaluations differed as a function of content (positive vs. negative). Once again, we did not obtain evidence to support the idea that evaluations differed as a function of information type (genuine vs. synthetic), in other words, evaluations established on the basis of Deepfaked audio were similar to those established via genuine audio.

In **Study 5** we wanted to replicate our previous findings (from Study 3 with videos) and also ask an additional set of questions - namely - are there specific demographic and/or individual difference factors that correlate with (a) one’s susceptibility for falling for Deepfaked videos and (b) the magnitude of their evaluations? In our previous studies a number of participants indicated (on the exploratory questions) that the videos were fake or manipulated in some sense whereas many others failed to discern this fact. We wanted to know what factors might help us predict whether one falls for a Deepfake attempt or is resilient to it. With this in mind, we replicated the same general design as in Study 3, with two notable changes. First, we utilized a different (algorithmic) process by which the Deepfaked videos were created in an attempt to improve their quality. Second, we included a number of demographic questions at the beginning of the study (age, gender, income, education, location, religious affiliation, social and political orientation) as well as individual difference questionnaires at the end of the study (preference for effortful vs. intuitive thinking [Rational-Experiential Inventory], cognitive ability [Cognitive Reflection Task]; overclaiming [overclaiming questionnaire]; and conspiratorial thinking [Beliefs in Conspiracy Theories]).

One the one hand, we set out to replicate (and thus confirm) our previous findings (i.e., that self-reported and automatic evaluations will vary as a function of information content [positive vs. negative information], and that this effect will not be moderated by information type [genuine vs. deepfaked]). On the other hand, we also set out to explore if (a) deepfake detection and (b) the magnitude of evaluative effects would vary as a function of the demographic and individual difference factors assessed in this study (e.g., will those who are lower on cognitive ability or preference for effortful thinking be more likely to endorse the Deepfaked videos as genuine, and/or show larger changes in liking; will those with more conspiratorial thinking or tendencies for overclaiming also show a similarly larger effect? In contrast, will those with higher levels of cognitive ability and/or preference for effortful thinking be more likely to recognize the Deepfaked videos as being manipulated, and/or show reduced evaluative learning effects relative to the aforementioned individuals)? Asking and answering such questions will allow us to not only confirm *that* Deepfakes are capable of altering evaluations in a similar way to genuine content, but also *who* is most susceptible to influence attempts via Deepfakes.