**Changing First Impressions via Deepfaked & Genuine Content**

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).

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)