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

In **Studies 1-6** we wanted to explore three questions concerning impression formation towards a novel individual. First, does first-hand information directly communicated by a target person lead to the formation of self-reported and automatic evaluations towards that individual? If so then the specific *content* of that information (positive or negative statements of their past actions) should influence the type of evaluations that are subsequently 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).

Second, we wanted to know if the *nature* of the information (synthetic vs. genuine) moderate 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. Third, would 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. In **Studies 5-6** a number of demographic and individual difference measures were included to explore this question. [[1]](#footnote-1)

In **Study 7** we wanted to carry out a high powered replication study designed to confirm the primary research questions of interest (for further details see the ‘Hypotheses’ document).

1. 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 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), (b) informational *type* (synthetic vs. genuine), or (c) *deepfake detection* serve to moderate self-reported and automatic evaluations as well as behavioral intentions? That said, 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)