**Data Analysis Plan**

**Data Preparation**

**Self-reported ratings**. A mean self-reported rating score will be calculated for Chris by averaging responses from the three Likert rating scales. Positive values indicate positive evaluations of Chris, negative values indicate negative evaluations, and neutral evaluations indicate neutral or ambivalent responses to Chris.

**IAT**. The D2 algorithm will be used to create pIAT scores. Scores will be calculated so that positive values reflected a relative response bias for Chris over Bob whereas negative values indicated the reverse pattern of responding (a relative response bias favoring Bob over Chris).

**Behavioral intentions.** A mean behavioral intentions score will be calculated for Chris by averaging responses from the three behavioral intention questions. Positive values agreement with the idea of supporting Chris’s YouTube channel, negative values indicate disagreement with the idea of supporting Chris’s YouTube channel, and neutral evaluations indicate neutral or ambivalent intentions to support Chris.

**Deepfake video detection**. Participants who are exposed to a Deepfake, and indicate “Yes” on the Deepfake detection question will be classified as having made an accurate judgement whereas those who indicate “No” on that question will be classified as having made an inaccurate judgement. Participants who are exposed to a genuine video and indicate “Yes” on the Deepfake detection question will be classified as having made an inaccurate judgement whereas those who indicate “No” on that question will be classified as having made an accurate judgement.

**Deepfake concept** **awareness**. Participants who indicate “Yes” on the Deepfake concept check question will be classified as having knowledge of Deepfakes prior to the study whereas those who indicate “No” will be classified as having no knowledge of Deepfakes prior to the study.

**Data Exclusions**

1. The data of participants who do not fully complete all questions and tasks will be excluded from analyses during the first round of analyses.

2. The data will be excluded of those participants who had pIAT error rates for any of the pIATs above 30% across the entire task, or above 40% for any one of the four critical blocks or for participants who complete more than 10% of pIAT trials faster than 400 ms.

**Hypothesis Testing**

**Confirmatory Analyses 1a: Genuine videos give rise to changes in evaluations and intentions.** Wewill examine this question in two ways. First, mean scores for the self-reported ratings, IATs, and behavioral intentions for those in the genuine video condition will be submitted to independent sample t-tests with *video content* (Positive vs. Negative) as a between subjects factor. This will tell us if the direction of evaluations differs depending on the video participants are exposed to, such that those in the positive video condition are expected to show positive self-reported ratings and IAT scores (as well as ambivalent intentions), whereas those in the negative video condition are expected to show negative self-reported ratings, IAT scores, and behavioral intention scores.

Second, we will submit mean scores for the self-reported ratings, IATs, and behavioral intentions to single sample t-tests to demonstrate that those scores differ from zero. Before we do so scores from those in the negative video condition will be multiplied by -1. In this way positive scores will indicate a change in attitudes/intentions in the expected direction, negative scores will indicate a change in an unexpected direction, whereas neutral values will indicate no change in attitudes/intentions (in such a case, neutral scores may indicate the absence of an evaluation or ambivalence).

In all cases, effect sizes (Cohen’s d) will be reported. We will also compute Bayesian factors in accordance with procedures outlined by Rouder, Speckman, Sun, Morey, and Iverson (2009) to estimate the amount of evidence for the hypothesis that stimulus evaluations differ as a function of video content (alternative hypothesis) or that there is no difference (null hypothesis).

**Confirmatory Analyses 1b: Deepfaked videos give rise to changes in evaluations and intentions.** Wewill examine this question in two ways. First, mean scores for the self-reported ratings, IATs, and behavioral intentions for those in the Deepfaked video condition will be submitted to independent sample t-tests with *video content* (Positive vs. Negative) as a between subjects factor. This will tell us if the direction of evaluations differs depending on the video participants are exposed to, such that those in the positive video condition are expected to show positive self-reported ratings and IAT scores and ambivalent intentions, whereas those in the negative video condition are expected to show negative self-reported ratings, IAT scores, and behavioral intention scores.

Second, we will submit mean scores for the self-reported ratings, IATs, and behavioral intentions to single sample t-tests to demonstrate that those scores differ from zero. Before we do so scores from those in the negative video condition will be multiplied by -1. In this way positive scores will indicate a change in attitudes/intentions in the expected direction, negative scores will indicate a change in an unexpected direction, whereas neutral values will indicate no change in attitudes/intentions (in such a case, neutral scores may indicate the absence of an evaluation or ambivalence).

**Confirmatory Analysis 2**: **Evaluations and intentions produced by the Genuine videos will not differ from those produced by the Deepfaked videos**. Mean self-reported ratings, IAT scores, and behavioral intentions will be submitted to independent samples t-tests to examine if the genuine and Deepfaked videos differ in the evaluations and intentions they produce. Data will first be recoded so that the valence of the video content is controlled for (i.e., scores from those in the negative content groups will be re-coded by multiplying their values by -1). Effect sizes (Cohen’s d) will be reported. We will also compute Bayesian factors in accordance with procedures outlined by Rouder, Speckman, Sun, Morey, and Iverson (2009) to estimate the amount of evidence that stimulus evaluations differ as a function of video type (alternative hypothesis) or that there is no difference (null hypothesis).

**Confirmatory Analysis 3a: 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.**

**Confirmatory Analysis 4: Changes in evaluations and intentions for participants who detect prior exposure to a Deepfaked video**. We will first select that subset of data from participants who were (a) exposed to a Deepfaked video and who (b) were classified as having detected the Deepfake video upon questioning. We will then carry out a similar set of analyses as outlined in Confirmatory Analysis 2 to determine if changes in evaluations and intentions still occur in this sample.