**Meta-Analysis (Experiments 1-5)**

A number of research questions/hypotheses were generated from exploration of the data from Experiments 1-5 that were not contained in the original preregistration for individual studies. Separately, some methodological improvements were generated after Experiments 1-5 were run (e.g., improved exclusion criteria to ensure participants stayed on the page where they watched/listened to the intervention in its entirety). We therefore elected to use the data from Experiments 1-5 to create a (non-preregistered) alternative analytic strategy (i.e., Bayesian multilevel models for each dependant variable) that formalized our core research questions, hypotheses, analytic models, inference rules, and other researcher degrees of freedom. This analytic strategy (and code to implement it) is described below, and was used in this meta-analysis, and later preregistered for Experiment 6, which was designed to provide strong confirmatory tests of these hypotheses.

For each hypothesis below, we specified how each verbal hypothesis corresponded to a statistical inference rule that would be used to conclude support for that hypothesis. We also report results from the exploratory analyses applied to Experiments 1-5 – this analytic strategy was developed on the existing data and was then preregistered and applied to Experiment 6 (with some necessary modifications, i.e., removing the random effect for experiment). The development of this precision in the implementation and interpretation of the analyses served to strengthen the later confirmatory analyses in Experiment 6.

All evaluative dependent variables (self-reported evaluations, IAT D2 scores, and behavioral intentions) were standardized (by 1 SD) after exclusions and prior to analysis condition (see Lorah, 2018: https://doi.org/10.1186/s40536-018-0061-2). This was done within each level of both IV (i.e., by Source Valence condition [positive vs. negative], and by Content [Genuine vs. Deepfaked]). As such, the beta estimates obtained from the Bayesian linear models (see research questions and data analysis plans below) therefore represent standardized beta values. More importantly, the nature of this standardization makes these estimates somewhat comparable to the frequentist standardized effect size metric Cohen’s *d*, as both are a difference in (estimated) means as a proportion of SD - although they should not be treated as equivalent. Effect size magnitude here can be thought of as using comparable scales as Cohen’s *d*. As such, to aid interpretability, the point estimates of these beta estimates are reported as δ (delta) rather than β.

**Exclusions**

In addition to the preregistered exclusion criteria (i.e., incomplete data or failure to maintain IAT performance criteria), participants were excluded if they spent too little or too much time viewing the web page that played the video or audio content, which may indicate that they did not watch or listen to the content or did not pay sufficient attention to it. We employed a minimum page linger time of 1.5 minutes and a max of 4.5 minutes on the basis that the intervention lengths varied between experiments and our goal was to exclude clear outliers and implausible values.

**Data Processing**

Our previous studies employed different variants of the IAT D score to score the pIAT data (Greenwald et al., 2003). For meta-analysis, all data was scored using the D2 variant.

**Analytic Strategy**

***Bayesian Models***

**Model Specification*.*** Bayesian models were implemented using the R package brms (Buerkner, 2017), which leverages the STAN language to allow for Bayesian inference via MCMC sampling.

**Linear Models***.* The linear models (hypotheses 1, 2, 5, 6, 7) took the following generic format: a dependent variable (IAT D2 score, self-reported ratings, or behavioral intentions); two dependent variables, *Source Valence* (positive vs. negative valenced statements) and *Content Type* (genuine vs. Deepfaked); and their interaction. When these were applied to the existing data from Experiments 1-5, a random intercept for Experiment was also added to the model (i.e., these were meta-analytic models).

E.g., Wilkinson notation for exploratory analyses of Experiments 1-5:

dependent\_variable ~ source\_valence \* content\_type + (1 | experiment)

*Poisson model.* The Poisson model (hypothesis 4) took the following format: cell counts served as dependent variable; two dependent variables, Deepfake concept awareness and Deepfake detection; and their interaction. When these were applied to the existing data (Experiments 4-5), a random intercept for Experiment was also added to the model (i.e., these were meta-analytic models).

E.g., Wilkinson notation for exploratory analyses of Experiments 1-5:

counts ~ awareness \* detection + (1 | experiment)

**Model Priors and their Informativeness.** Wide priors have been specified for all parameters (i.e., normal distribution with *M* = 0 and *SD* = 10, following general recommendations for weakly informative priors in STAN: <https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations>). We used Gelman’s (2019) method to characterize in order to characterize the priors as uninformative: For each parameter, we compared the posterior SD to the prior SD. If the posterior SD for any parameter was more than 0.1 times the prior SD, we noted that the prior was informative, otherwise it was noted as uninformative. Inspection of prior and posterior distributions for the models fit to the data from our previous experiments (1-5) allowed us to conclude that all priors were uninformative. As such, results (i.e., derived from posterior distributions) were very weakly influenced by the prior, and therefore likely to be comparable to what would be found had we used frequentist estimation methods (i.e., driven in large part by the data rather than the prior).

**Model Convergence*.*** We inspected the convergence of the chains via visual inspection of the plots, , and the effective sample size metrics. Appropriate changes to model hyper parameters were made if evidence of non-convergence was found (e.g., increasing number of iterations or the adapt\_delta parameter and refitting the model).

**Parameter Estimation and Inference*.*** Posterior distributions were summarized via a metric of central tendency, the Highest Maximum A Posteriori probability estimate (MAP). This was judged to be a preferable metric to the mean given the mean’s sensitivity to outliers. Estimation width was quantified via 95% Credible Intervals via asymmetric Highest Density Intervals (HDIs). In the linear models, estimates for subgroups were calculated via manipulation of the posterior probabilities (e.g., genuine condition = intercept, Deepfaked condition = intercept + main effect for experiment condition, etc.; see R code implementation for details).

Bayesian *p* values were also produced for the sake of familiarity for many readers. These were derived from the proportion of the posterior samples that were in the predicted direction: Bayesian *p* = ≈ frequentist *p* value (where refers to = 0 in the linear models or IRR = 1 in the Poisson model). All three of these metrics were implemented using the bayestestR R package.

**Null-Hypothesis Test*.*** Null-hypothesis tests (e.g., for H1, H4, and H5) were implemented via the inspection of the 95% Credible Intervals. If a CI’s lower bound was > (where refers to = 0 in the linear models or IRR = 1 in the Poisson model), this was considered evidence in support of the alternative hypothesis (e.g., that the estimated means differed).

**Non-Inferiority Tests*.*** Non-inferiority tests (e.g., for H2) were implemented via the general method described by Lakens, Scheel, and Isager (2018), albeit (1) applied to intervals derived from Bayesian models and (2) applied unidirectionally (i.e., as a non-inferiority rather than equivalence test). Specifically, if the lower bound of the 95% CI of the genuine condition was < the lower bound of the 90% CI of the Deepfaked condition (i.e., the difference between Source Valence conditions in each subgroup), this was considered evidence in support of the alternative hypothesis (i.e., evidence of non-inferiority in estimated means; that Deepfakes are as good as genuine content).

In addition to this non-inferiority hypothesis test, which we note is a relatively strict test, an effect size was produced to characterize the magnitude of the effect size in the Deepfaked condition as a percentage of the genuine condition. This was implemented by calculating a proportion for each posterior sample and then parameterizing this new distribution (via MAP and 95% HDI). In addition to the above non-inferiority test, we concluded that Deepfaked content produces substantively similar effect impression formation (in a continuous rather than categorical sense) by describing this estimate of comparative effect size (e.g., that the magnitude of the Deepfake condition was within ± 10% of genuine content).

**Classification Statistics.** Many have argued that no single classification metric is optimal. Therefore a confusion matrix and multiple classification metrics were calculated using the true status of the video content (genuine or Deepfaked) and participants Deepfake detection responses, specifically: False Positive Rate, False Negative Rate, Balanced Accuracy, and Informedness (Youden’s *J*). 95% Confidence Intervals were bootstrapped using the case removal and percentile methods and 2000 iterations.

**Hypothesis Testing**

**Research Question 1: Does Online Content Change Attitudes and Intentions Towards a Novel Individual?**

We first wanted to know if, *in general*, the informational content of genuine and Deepfaked content influenced people’s attitudes and intentions. We tested this using a Bayesian linear model. Doing so allowed us to estimate a 95% Credible Interval on standardized effect size change in evaluations between Source Valence conditions (i.e., between those who encountered the positive or negative variant of the content). Credible Intervals whose lower bounds were > 0 were viewed as support for a given hypothesis. We explored this for each outcome measure.

Meta-analytic results indicated that the informational content of genuine videos (i.e., Source Valence) influenced self-reported attitudes (Standardized effect size *δ =* 2.71, 95% CI [2.57, 2.85], *p* < .0000001), implicit attitudes (*δ =* 1.33, 95% CI [1.19, 1.46], *p* < .0000001), and behavioral intentions (*δ =* 1.13, 95% CI [0.73, 1.52], *p* < .0000001). The same was true for Deepfaked content, which also influenced self-reported attitudes (*δ =* 2.71, 95% CI [2.53, 2.88], *p* < .0000001), implicit attitudes (*δ =* 1.32, 95% CI [1.16, 1.49], *p* < .0000001), and behavioral intentions (*δ* = 3.06, 95% CI [2.67, 3.45], *p* < .0000001).

**Research Question 2: Are Deepfakes as Effective as Genuine Content at Influencing Attitudes and Intentions?**

We then examined if Deepfaked content was as effective (i.e., non-inferior) to genuine content when it came to changing attitudes and intentions. If the lower bound of the 95% CI of the genuine condition was < the lower bound of the 90% CI of the Deepfaked condition (i.e., the difference between Source Valence conditions in each subgroups), we considered this as evidence in support of the alternative hypothesis (i.e., evidence of non-inferiority in estimated means; that Deepfakes are as good as genuine content). In addition to the relatively strict non-inferiority test, we also compared the magnitudes of the effect sizes to make more general comparisons about their comparative effectiveness (e.g., to observe that the magnitude of the Deepfake condition was within ± 10% of genuine content).

Meta-analyses revealed that self-reported attitudes induced by Deepfaked content were non-inferior to genuine content (genuine lower 95% CI = 2.57; Deepfake lower 90% CI = 2.57). Deepfakes were 100.4% (95% CI [92.1, 108.8]) as effective in changing self-reported attitudes as their genuine counterparts. A similar pattern emerged for implicit attitudes (pIAT scores): Deepfaked content was non-inferior to genuine content here too (genuine lower 95% CI = 1.19; Deepfake lower 90% CI = 1.18). Deepfaked content was 100.3% (95% CI [83.9, 117.1]) as effective in changing implicit attitudes as genuine content. Finally, behavioural intentions induced by Deepfaked content were also non-inferior to genuine content (genuine lower 95% CI = 0.73; Deepfake lower 90% CI = 2.75). Deepfakes were 259.5% (95% CI [184.8, 405.3]) as effective in changing intentions as genuine content.

**Research Question 3: How Effective are People at Detecting Deepfakes?**

In Experiments 3-5, participants were first told what a Deepfaked was, informed that they had been exposed to one, and asked to indicate in an open-ended response whether they had been aware of this fact while watching the content (i.e., if they were aware that the content was Deepfaked while watching it). These open-ended responses were then coded as “Yes” or “No” by two independent raters. If both raters scored a response as having classified the content as a Deepfake then it was scored as such, otherwise they were scored as genuine (i.e., scoring prioritized specificity over sensitivity). Good agreement was found between raters (92% agreement, Cohen’s = .78, 95% [.72, .84]).

We examine if participants could make accurate and informed judgements about whether online content was genuine or Deepfaked. Analyses revealed that manyparticipants incorrectly believed that the Deepfake was actually a genuine video (false negative rate = .73, 95% CI [.69, 0.78]), and that a small number incorrectly believed that the genuine content was Deepfaked (false positive rate = .08, 95% CI [.04, 0.12]). We also found that participants were poor at making accurate decisions about whether content is genuine or not (e.g., Balanced Accuracy = .59, 95% CI [.56, 0.62]), and poorly informed decisions about whether content is genuine or not (e.g., informedness/Youden’s *J* = .19, 95% CI [.13, .25]).

**Research Question 4: Are People Aware That Content Can Be Deepfaked Before They Take Part in The Study and Does This Make Them Better at Detecting Them?**

In Experiments 4-5, we asked participants if, prior to the study, they knew that video or audio content could be Deepfaked (i.e., if they were aware of the general concept of Deepfakes). They provided their responses in an open-ended fashion, and these responses were then coded as “Yes” or “No” by two other independent raters. If both raters scored a response as having classified the content as Deepfake aware then it was scored as such, otherwise they were scored unaware. Inter-rater reliability was found to be good. Results indicated that roughly half (53.5%) of participants were scored as aware of the concept of Deepfakes prior to the study.

We then examined if participants who reported being aware of Deepfaking prior to the study would also be better at detecting Deepfakes when exposed to one. Specifically, using the subset of participants who were in the Deepfake condition, we calculated counts for each of the combinations of the Deepfake concept check and Deepfake detection questions (e.g., awareness = TRUE & detection = TRUE, awareness = TRUE & detection = FALSE, etc.). We then used a Bayesian Poisson model to estimate a 95% Credible Interval around the Incidence Rate Ratio. A Credible Interval whose lower bound is > 1 was considered evidence in support of this hypothesis. Estimated marginal predicted probabilities are also reported.

Results indicated that participants who were aware of Deepfaking were also twice as likely to correctly detect a Deepfake when they were exposed to one (IRR = 2.75, 95% CI [1.41, 5.35]). Specifically, those who were previously unaware of Deepfaking had a 6% chance of detecting it whereas their aware counterparts had a 14% chance of detecting it.

**Research Question 5: Does Prior Awareness of the Concept of Deepfakes Make You Immune to Their Influence?**

We examined if attitudes and intentions would still emerge for ‘aware’ participants (i.e., those who were exposed to a Deepfake and who reported being aware of the concept of Deepfaking prior to taking part). Results indicated that prior awareness of Deepfaking did not protect an individual from being influenced by the Deepfake. Aware individuals also showed changes in self-reported attitudes, *δ =* 3.08, 95% CI [2.66, 3.48], *p* < .0000001, implicit attitudes, *δ =* 1.40, 95% CI [1.01, 1.74], *p* < .0000001, and behavioral intentions, *δ =* 3.09, 95% CI [2.52, 3.67], *p* < .0000001.

**Research Question 6: Does Detecting Deepfaked Content Protect One From Its Influence?**

We also examined if participants who successfully detected the presence of a Deepfake would also be immune to its influence. Deepfake detectors were also influenced by such content, and showed a change in self-reported attitudes, *δ* = 2.72, 95% CI [2.30, 3.23], *p* < .0000001, implicit attitudes, *δ* = 1.06, 95% CI [0.69, 1.42], *p* < .0000001, andbehavioral intentions,*δ* = 2.70, 95% CI [1.91, 3.55], *p* < .0000001.

**Research Question 7: Does Awareness and Detection of Deepfakes Protect One from Its Influence?**

Finally, we wanted to know if individuals who were both aware of Deepfaking prior to the study *and* who successfully detected the presence of the Deepfake, would be immune to the Deepfakes influence. Results indicated that both awareness and Deepfake detection did not immunize the individual from its influence, such that these participants also showed the expected change in self-reported attitudes, *δ* = 3.28, 95% CI [2.32, 4.23], *p* < .0000001, implicit attitudes, *δ* = 1.23, 95% CI [0.58, 1.91], *p* < .0000001, and behavioral intentions,*δ* = 2.48, 95% CI [1.42, 3.57], *p* < .0000001.