Computers in Human Behavior 149 (2023) 107917

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Providing detection strategies to improve human detection of deepfakes: An experimental study

Klaire Somoray [*], Dan J. Miller

Department of Psychology, College of Health Care Sciences, James Cook University, Australia

ARTICLE INFO

Handling Editor: Jen-Her Wu

Keywords:
Deepfake
Human detection
Deepfake detection strategies
Accuracy
Confidence

ABSTRACT

Deepfake videos are becoming more pervasive. In this preregistered online experiment, participants ($N = 454$, $M_{\text{age}} = 37.19$, $SD_{\text{age}} = 13.25$, males = 57.5%) categorize a series of 20 videos as either real or deepfake. All participants saw 10 real and 10 deepfake videos. Participants were randomly assigned to receive a list of strategies for detecting deepfakes based on visual cues (e.g., looking for common artifacts such as skin smoothness) or to act as a control group. Participants were also asked how confident they were that they categorized each video correctly (per video confidence) and to estimate how many videos they correctly categorized out of 20 (overall confidence). The sample performed above chance on the detection activity, correctly categorizing 60.70% of videos on average ($SD = 13.00$). The detection strategies intervention did not impact detection accuracy or detection confidence, with the intervention and control groups performing similarly on the detection activity and showing similar levels of confidence. Inconsistent with previous research, the study did not find that participants had a bias toward categorizing videos as real. Participants overestimated their ability to detect deepfakes at the individual video level. However, they tended to underestimate their abilities on the overall confidence question.

1. Introduction

“Deepfake” is an umbrella term for a wide variety of computer-generated synthetic media, in which a person in an existing image or video is manipulated to have another person’s likeness. This technology usually results in highly realistic media. The application of machine learning methods to alter video footage began in the late 1990s (Bregler et al., 1997). However, general adversarial networks—the machine learning technique used to create deepfakes—weren’t developed until the 2010s (Mirsky & Lee, 2021). Deepfakes came to the attention of the public in 2017 after a Reddit user named “deepfakes” shared videos they had created to the website, resulting in a hobbyist community centred around the subreddit r/deepfakes (Cole, 2018; Mirsky & Lee, 2021). Other Reddit users later created software like FakeApp, which enabled the creation of deepfakes with minimal programming experience (Cole, 2018; Mirsky & Lee, 2021). r/deepfakes primarily produced humorous or pornographic deepfakes of celebrities (Westerlund, 2019) and was eventually shut down by Reddit (Doctorow, 2018).

The ease with which high-quality deepfakes can now be generated has raised concerns about this technology being used for nefarious purposes beyond the creation of deepfake pornography. For instance, deepfakes of prominent figures have already been used to spread political misinformation. In March 2022, a manipulated video of the Ukrainian President, Volodymyr Zelensky, was circulated. In this video, Zelensky is depicted appealing to Ukrainian soldiers to surrender. Although the video itself was easily detected as fake and mostly ridiculed (see Wakefield, 2022), deepfake technology is increasing in sophistication. Others have raised the possibility of deepfake technology being used to create white supremacist propaganda (Habgood-Coote, 2023). Concerns regarding this technology have driven research into the public’s ability to detect deepfakes (e.g., Groh et al., 2022; Kobis et al., 2021; Korshunov & Marcel, 2020).

1.1. People’s ability to detect deepfakes

The majority of deepfake detection studies have focused on artificial intelligence for classifying videos as real or fake. Several review papers have looked at the efficacy of these technologies (e.g., Passos et al., 2022; Tolosana et al., 2020). These reviews found that deepfake detection by machine models can range from around 60%–100%,

[*] Corresponding author.
E-mail address: klaire.somoray@jcu.edu.au (K. Somoray).

https://doi.org/10.1016/j.chb.2023.107917
Received 28 February 2023; Received in revised form 16 June 2023; Accepted 13 August 2023
Available online 21 August 2023

although accuracy rates are highly dependent on many factors in addition to the model used (e.g., dataset used). While AI technologies for deepfake detection exist, their implementation in social media platforms (where these videos are being shared) is yet to be seen. Furthermore, these deepfake detection technologies are not accessible to the public. Thus, humans are often left on their own to decide whether videos are fake or authentic.

Despite the potential cognitive biases of humans, some researchers strongly believe in the "wisdom of the crowd" in relation to deepfake detection (Groh et al., 2021, 2022). While it is important to acknowledge that state-of-the-art machine learning techniques can have a detection performance of up to 100%, humans can outperform some computer models in deepfake detection (Groh et al., 2021, 2022). Groh et al. (2021) attributed this to humans' specialized ability in visually processing faces. Deepfakes are computer-generated synthetic media and some artifacts are produced by deepfake-generator algorithms. These artifacts, which may not be salient enough for computer vision, may be more perceptible to humans due to our ability to process faces holistically.

Similarly to computer models however, human performance on deepfake detection also varies, with studies presenting accuracy levels ranging from 57.6% (Kobis et al., 2021) to 88.9% (Groh et al., 2022). It is difficult to compare these accuracy ratings, given methodological differences in the datasets used and the way accuracy scores are calculated. In Groh et al.'s (2022) second experiment, participants' accuracy scores were indexed as a function of whether participants correctly categorized videos as deepfake or real *and* their level of confidence in this categorization. For example, if a participant correctly detected a deepfake with 82% confidence in their categorization, this participant was assigned an accuracy score of 0.82. If the categorization was incorrect, the participant would be assigned an accuracy score of 0.18. Other studies (e.g., Kobis et al., 2021; Rossler et al., 2019) only assessed whether participants correctly classified videos as real or deepfake (using a dichotomous forced-choice format). Korshunov and Marcel (2020) took a similar approach but also provided a third "I do not know" option for participants.

1.2. Biases in human detection of deepfakes

Prior research indicates that participants tend to have a bias toward categorizing videos as authentic (e.g., Kobis et al., 2021; Korshunov & Marcel, 2020). Korshunov and Marcel (2020) found that "good quality" deepfake videos can "easily" fool the public, with only 24.5% of good quality deepfakes being perceived as fake. Similarly, participants in Kobis et al.'s study (2021) categorized videos as real 67.4% of the time, even though they were explicitly told that only half of these videos were real. Kobis et al. (2021) suggest that this bias could be attributed to an overly optimistic "seeing-is-believing" heuristic. In other words, people will conclude that video content is authentic until there is clear evidence otherwise. In the current study, we expect to find the same effect.

In addition to being poor detectors, people also tend to overestimate their ability to detect deepfakes. In Korshunov and Marcel's (2020) study, very few participants indicated uncertainty (i.e., choosing the "I don't know" option), suggesting that people seemed to be sure when judging the realism of deepfake videos. Kobis et al. (2021) found that participants significantly overestimated the number of videos they correctly categorized. They also observed a negative correlation between overconfidence and accuracy. The authors attributed these findings to the Dunning-Kruger effect. The Dunning-Kruger effect is a cognitive bias in which those who lack competence in a domain tend to be ignorant of their own incompetence within this domain (precisely because of their lack of knowledge), resulting in low performers overestimating their own ability (Dunning, 2011; Kruger & Dunning, 1999). The existence of this effect remains controversial, with some suggesting that the effect may just be a "data artifact" (Ackerman et al., 2002; Gignac & Zajenkowski, 2020; Nuhfer et al., 2017).

1.3. Public intervention for deepfake detection: what helps?

To date, very few studies have assessed the efficacy of interventions to improve humans' detection of deepfakes. For example, Groh et al. (2022) looked at the impact of providing human participants with AI predictions (as to whether a video is fake or real) on accuracy. This intervention significantly increased participants' rate of correct identification (from 66% to 73% of observations). Interestingly, Groh et al. (2022) also found that longer time to complete the identification activity was associated with poorer detection accuracy.

In another study, Kobis et al. (2021) investigated whether increasing participants' motivation to correctly detect deepfake videos (by giving a financial incentive for correct categorizations or by raising awareness of the negative societal consequences of deepfakes) impacts detection accuracy. However, these interventions did not have a significant effect, with participants in intervention groups performing similarly to the control group.

Several organizations have developed guidelines to help the public identify deepfake media (eSafety, 2022; MIT Media Lab, n. d.). For instance, researchers from MIT have created a list of deepfake artifacts, such as facial transformations, lighting, smoothness of cheeks and forehead (MIT Media Lab, n. d.), which can be applied when assessing the veracity of videos. Additionally, Australia's independent regulator for online safety, eSafety, recently published a position statement on deepfakes along with advice for identifying deepfake media (eSafety, 2022). This advice largely mirrors the strategies developed by MIT. However, the question remains as to whether these applying these strategies actually improve deepfake detection.

1.4. Current study

The main objective of the current study is to investigate whether a relatively simple intervention (based on providing strategies for detecting visual artifacts) is effective in helping the public to identify deepfakes. In our experiment, half of the participants were presented with strategies for detecting deepfakes, while the other half served as a control group. An example strategy includes: "Pay attention to the cheeks and forehead. Does the skin appear too smooth or too wrinkly? Is the agedness of the skin similar to the agedness of the hair and eyes? DeepFakes are often incongruent on some dimensions." Knowing whether this intervention is effective or not has the potential to benefit the wider community by shedding light on the mechanisms underlying human detection of deepfakes. We have the following hypotheses:

H1. Participants will be biased toward categorizing stimulus videos as real.

H2. Participants provided with detection strategies will show greater detection accuracy relative to a control group.

H3. Participants provided with detection strategies will show greater detection confidence relative to a control group.

In addition to the above hypotheses, we have developed the following research questions:

RQ1. Is detection confidence associated with detection accuracy?

RQ2. Is level of interaction with stimulus videos is associated with detection accuracy?

2. Method

2.1. Design

This online study employed a between-subjects experimental design in which participants were randomly assigned to receive strategies for detecting deepfakes or not (control condition), before being asked to categorize a series of videos as fake or authentic. The study was hosted via Qualtrics and Qualtrics's randomizer function was utilized to randomly assign participants to conditions. All hypotheses, research questions, methodology and measures were preregistered on May 6, 2022, prior to the start of data collection (end of May 2022). Details of the preregistration can be found at https://osf.io/xdhck/.

A priori power analysis was conducted to determine the total sample size needed to test the efficacy of the intervention with a power of .80, assuming a small-to-medium effect ($d = 0.35$). This analysis indicated that an N of 204 would be sufficient (if using an α of 0.05 and performing a one-tailed tests).

2.2. Participants

The final sample ($N = 454$) had a mean age of 37.18 years ($SD = 13.25$). Other demographic characteristics of the sample are reported in Table 1.

2.3. Materials

2.3.1. Stimulus videos

Stimulus videos were sourced from the open-source *DeepFake Detection Challenge* (DFDC) dataset (https://arxiv.org/abs/2006.07397). The DFDC is a large database of videos from paid performers talking about various topics in different settings. The DFDC dataset was chosen as using well-known deepfakes of politicians and/or celebrities has the disadvantage of potential biases, such as prior exposure or emotional motivation to believe or discredit the video. Previous studies on human deepfake detection have used this dataset (Groh et al., 2022; Kobis et al., 2021; Korshunov & Marcel, 2020). The videos were selected at random until we had 20 pairs of videos (an authentic and deepfake version of each video) which met our inclusion criteria (consistent lighting with only one person depicted in the video). The selected deepfake videos were categorized as *difficult* to *very difficult* to identify as deepfakes (as categorized by Korshunov & Marcel, 2020). Ten videos depicted males and ten depicted females. The stimulus videos depicted models of various races. All videos were 10 s in length.

Following Kobis et al. (2021), stimulus videos were divided into two sets, with each participant only seeing one of the two sets (see Fig. 1). Stimulus videos were presented one at a time. The order of the presentation of videos was randomized. Furthermore, the presented stimulus videos had an equal number of female and male actors, as well as dark and light skinned individuals. All videos had one actor only. No actors wore glasses. The lighting conditions across the videos were consistent (except for two videos where the lighting is a bit dark) and none of the videos were blurry or grainy.

Table 1
Participant demographics and device used when completing study (N = 454).

	Frequency	%
Gender		
Male	261	57.5
Female	175	38.5
Non-binary	10	2.2
Missing	8	1.8
Country of Residence		
Australia	315	69
Canada	45	10
New Zealand	23	5
UK	30	7
United States	22	5
Other European country	10	2
Other, outside of Europe	4	1
Missing	5	1
Highest Level of Education		
Primary school	8	1.8
High school graduate	67	14.8
TAFE/Other vocational studies	83	18.3
Undergraduate degree	162	35.7
Some postgraduate study	130	28.6
Missing	4	0.9
Device Used		
Smart phone/Tablet (inc. iPad)	189	41.6
PC	265	58.4

2.3.2. Detection strategies

Those in the detection strategies condition were provided with 10 strategies for detecting deepfakes. These strategies were sourced from the MIT Media Lab (https://www.media.mit.edu/projects/detect-fakes/overview/). An example strategy is "Pay attention to the facial hair or lack thereof. Does this facial hair look real? Deepfakes might add or remove a moustache, sideburns, or beard. But, deepfakes often fail to make facial hair transformations fully natural."

2.3.3. Detection Accuracy and Confidence measures and demographics

After each stimulus video, participants were asked "Is this video a deepfake or real?" (binary response option: *This video is a deepfake*; *This video is real*) and "What is your confidence that you guessed correctly?" (measured using an unnumbered graphic rating scale anchored by *50 = As confident as flipping a coin* and *100 = 100% sure*). Detection accuracy was calculated by adding the number of videos correctly categorized and then dividing by the total number of videos categorized. Average detection confidence was calculated by averaging confidence responses across the 20 videos.

After the detection activity, participants' overall detection confidence was measured via a single item: "Out of 20 videos, how many videos do you think you guessed correctly?" This value was divided by 20 to make the scoring format consistent with the scoring format for average detection confidence. Participants were also asked about basic demographics including age, gender, country of residence, and highest level of formal education. To facilitate the test of H1, the percentage of videos that participants categorized as being authentic was also calculated.

2.3.4. Metadata

Qualtrics recorded several pieces of metadata, including time spent on each stimulus video page, number of clicks on each stimulus video page, and operating system used to access the study. Time spent and number of clicks was averaged across the 20 stimulus videos.

2.4. Participant recruitment

Participants were recruited via social media. Specifically, the study was posted to the subreddits r/SampleSize and r/Australia. A small, paid advertising campaign was also conducted on Facebook and Instagram. This involved creating a post about the study on a dedicated lab group Facebook page and then "boosting" this post between May 27, 2022, and May 31, 2022. Adults living in Australia, Canada, New Zealand, the United Kingdom and the United States were selected as the demographic group to be targeted. Based on URL referrer information collected by Qualtrics, one-third of participants were recruited via Reddit (35.2%) and 31.9% were recruited via Facebook. The remaining participants were missing this information (31.9%) or accessed the study from another platform (0.9%). Participants were not offered an incentive for their participation.

2.5. Procedure

Ethical approval to conduct the study was granted by the Human Research Ethics Committee of James Cook University (ID# H8747).

After informed consent was collected, participants were told that: 1) they will be shown 20 videos, 2) exactly half of these videos will be deepfakes, 3) they can watch these videos as many times as they need, and 4) they will be given a score indicating how many videos they correctly categorized. Based on Kobis et al. (2021), participants were then presented with two validity check items: "A deepfake is a video in

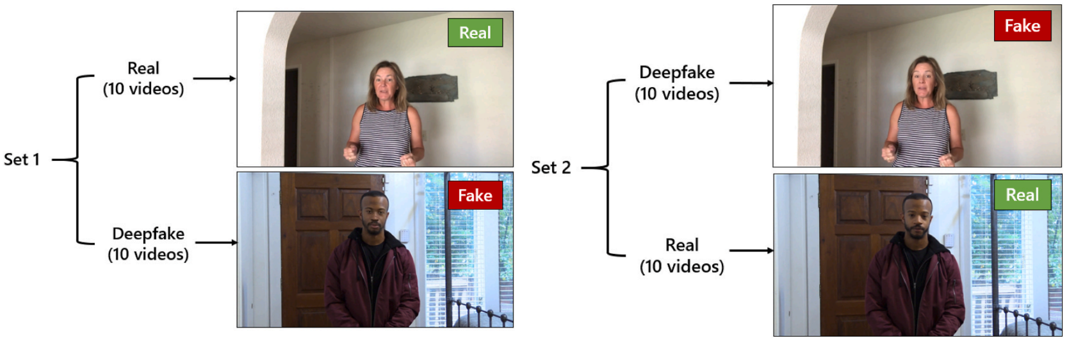

Fig. 1. Illustration of the two sets of videos presented to participants.

which …" (*A person is telling lies*; *A person is talking about fakes*; and *The face and/or voice of a person has been manipulated using artificial intelligence*) and "Which of the following statements is correct?" (*Each video has a 20% chance to be a deepfake*; *Each video has a 50% chance to be a deepfake*; and *All the videos are deepfakes*). These items needed to be answered correctly before participants could move on to the detection activity. Participants were then presented with the list of detection strategies (if applicable), the detection activity, and the overall confidence and demographic measures. As an additional validity check, participants were also asked to indicate if they had done this study previously. Finally, participants were provided with their detection accuracy score and debriefing information.

2.6. Data cleaning and analysis approach

A total of 762 participants accessed the study. Of these participants, 4 did not provide consent and a further 184 did not complete any of the stimulus video questions and were removed. A further 108 participants were removed for missing more than 50% of responses. Three participants indicated they were doing the study for a second time and were also removed. Finally, 12 participants were excluded for spending an average of under 15 s on each stimulus video page (as it was reasoned that these participants were unlikely to have watched the stimulus videos in their entirety).[1] These exclusions left a final sample of 454 participants.

Univariate outliers were detected using the outlier labelling rule with a 2.2 multiplier (Hoaglin & Iglewicz, 1987). Univariate outlying values were replaced with the next lowest/highest non-outlying value observed in the dataset. Outlying datapoints were observed on the following variables: percentage of videos categorized as real (11 outlying datapoints); average time spent per stimulus video (15 outlying datapoints); and average number of page clicks per stimulus video (2 outlying datapoints). Normality of variables was assessed via visual inspection of histograms and with reference to skewness and kurtosis values. In all cases, variables were observed to be acceptably normally distributed for the use of parametric tests.

One-tailed tests were used to test preregistered hypotheses. Two-tailed tests were applied when testing preregistered research questions and for any exploratory analyses not included in the preregistration.

3. Results

3.1. Detection accuracy

As stated in our preregistered plan, we computed an overall accuracy score for each participant representing the percentage of videos which the participant correctly categorized. Mean accuracy was just over 60% ($M = 0.61$, $SD = 0.13$), indicating that, on average, participants correctly detected 12 out of 20 videos (we would expect participants to correctly categorize 10 out of 20 videos by chance alone). The poorest performers correctly categorized 5 out of 20 videos (0.25) and the best performers correctly categorized 19 out of 20 (0.95). Average accuracy exceeded chance levels as indicated by a one-sample *t*-test, $t\ (450) = 16.91$, $p < .001$, Cohen's $d = 0.80$. On average, participants took 35.0 s ($SD = 19.1$) on each page to respond to the videos.[2]

3.2. Are participants biased and overconfident?

We hypothesised that participants would have a bias toward categorizing stimulus videos as real (H1). This hypothesis was not supported. A one-sample *t*-test indicated that the percentage of videos categorized as real among the overall sample ($M = 0.51$, $SD = 0.11$) was not significantly different to the reference value of 0.50, $t\ (450) = 1.33$, $p = .092$, Cohen's $d = 0.06$. As an auxiliary analysis (not included in the study preregistration), this same test was conducted among each experimental group, with non-significant results being observed in both the control and intervention group.

We were also interested the relationship between confidence and accuracy (RQ1). Detection accuracy displayed a small positive relationship with both overall detection confidence, $r\ (450) = 0.18$, $p < .001$, and average detection confidence, $r\ (451) = 0.22$, $p < .001$. Plots depicting these associations are provided as supplementary figures (Figs. S1 and S2).

Interestingly, participants, on average, correctly identified approximately 12 out of 20 videos ($M = 0.61$, $SD = 0.13$). However, when asked "Out of the 20 videos, how many videos do you think you guessed correctly?", participants suggested around 10 out 20 videos on average ($M = 0.51$, $SD = 0.19$). When confidence scores were assessed per video, a different pattern emerged. Participants showed greater confidence at the individual video level ($M = 0.74$, $SD = 0.11$). As an auxiliary analysis, this difference in overall detection confidence and average detection confidence per video was further explored via a paired-samples *t*-test. This test indicated that average per video detection confidence was significantly greater than overall detection confidence $t\ (449) = 29.57$, $p < .001$, Cohen's $d = 1.40$. Additionally, accuracy scores were greater than overall detection confidence scores, $t\ (450) = 9.76$, $p < .001$, Cohen's $d = 0.46$. However, average per video detection confidence

[1] This diverges from the study preregistration: "Those who complete the study much faster than average (under half the sample's 5% trimmed mean time to complete) will be excluded." However, we felt that this approaches better targeted our validity concern (i.e., that participants were not watching the stimulus videos).

[2] After the removal of outliers.

scores were found to exceed accuracy scores, t (453) = 18.41, $p < .001$, Cohen's d = 0.86, suggesting overconfidence (these analyses were not included in the preregistration). As can be seen in Fig. 2, participant confidence greatly exceeded accuracy across the majority of videos.

3.3. *Did the detection strategies help?*

Tests of H2 and H3 are presented in Table 2. As can be seen, neither hypothesis was supported. That is, the control and detection strategies groups did not differ on detection accuracy or detection confidence (overall confidence or average confidence per video). Fig. 3 shows the distribution of scores in the control and detection strategies groups for detection accuracy (Fig. 3A) and detection confidence (Fig. 3B and C).

Following a similar analysis by Kobis et al. detection accuracy by video type was assessed to gain a deeper insight into how participants responded (see Fig. 4). Interestingly, the same number of deepfake and authentic videos (n = 4) failed to reach the chance levels of accuracy.

With reference to RQ2, level of interaction with stimulus videos (as indexed by average time spent per stimulus video) was found to be unrelated to detection accuracy, r (451) = −0.03, p = .588. As an auxiliary analysis (not included in the preregistration), the average number of page clicks per stimulus video was calculated as an additional index of level of interaction with stimulus videos (as those who clicked stimulus video pages multiple times would be more likely to have watched the stimulus videos multiple times). Similarly to time spent per video, the average number of page clicks was found to be unrelated to detection accuracy, r (451) = −0.03, p = .474. Additional analyses were conducted for variables that may have had an impact on detection accuracy (video position, actor action, actor gender, actor skin color, device and operating system used to complete the study). These analyses are reported in the supplementary materials (Figs. S3 and S4).

Table 2
Analyses comparing treatment groups on detection accuracy and confidence.

Variable	*M (SD)*		*t*	*df*	*p*	*d*
	Control	Detection Strategies				
Detection accuracy	.60 (.14)	.61 (.13)	−1.20	449	.115	−0.11
Overall detection confidence	.51 (.19)	.51 (.18)	−0.61	449	.270	−0.06
Average detection confidence	.74 (.10)	.74 (.11)	−0.18	448	.427	−0.02

Note. All tests reported in table are one-tailed.

4. Discussion

4.1. *Participant accuracy in detecting deepfakes*

This study adds to a growing literature examining people's ability to detect deepfakes. Our participants accurately detected deepfakes, on average, 60.7% of the time (statistically significantly above chance). This detection score is slightly higher than that observed in Kobis et al.'s (2021) study: 57.6%. Contrary to our hypothesis, participants did not display a bias toward categorizing stimulus videos as real. This also contrasts with the findings of Kobis et al.'s (2021), where participants suggested that videos were real 67.4% of the time (even though they were informed that only half of the videos were authentic). Kobis et al. (2021) attributed this to a "seeing-is-believing" heuristic in which videos are assumed to be real unless evidence clearly indicates otherwise.

An explanation for this inconsistency could be due to difference in recruitment strategies between studies. Participants in Kobis et al.'s (2021) study were recruited from Profilic (a participant recruitment website) and received a participation payment of £2.5. In our study,

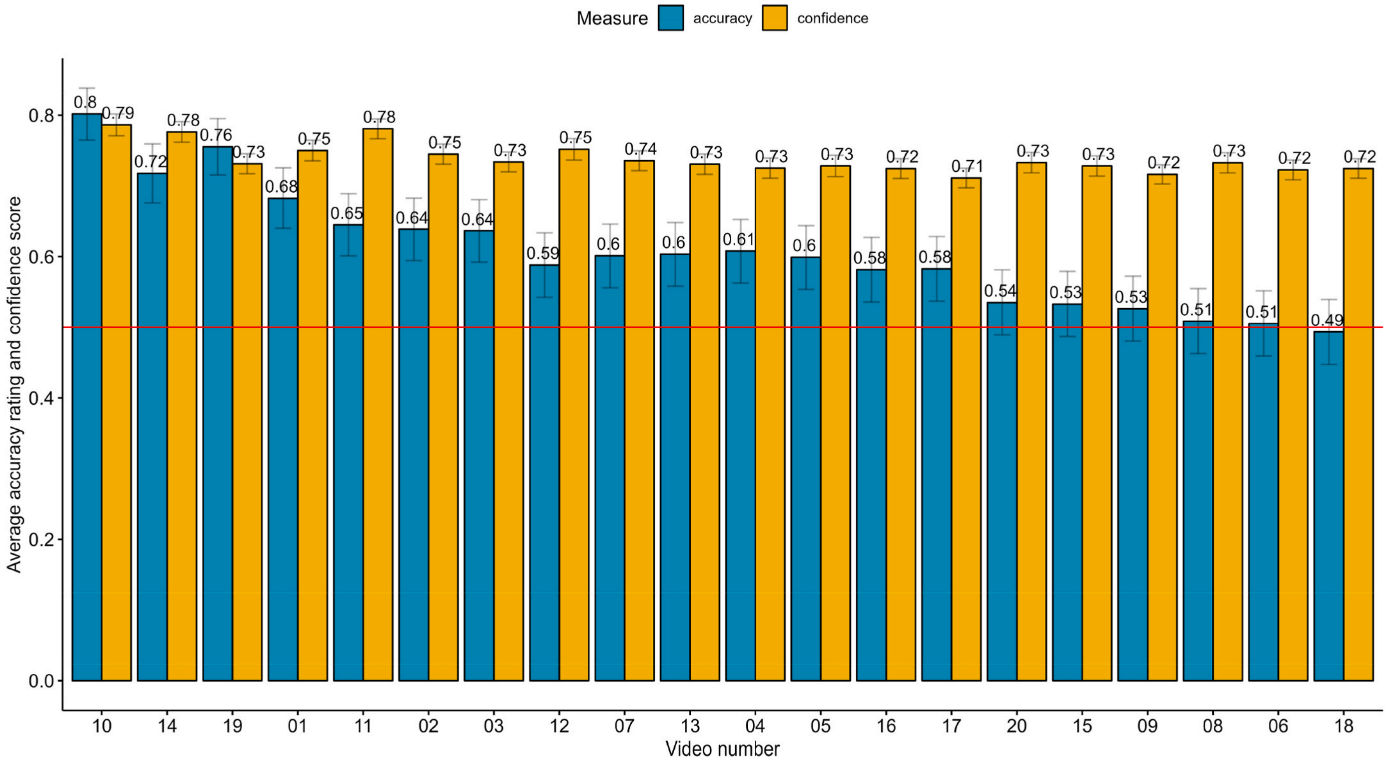

Fig. 2. Accuracy and Confidence by Video
Note. The numbers presented in the plots are the average accuracy scores (blue) and average confidence scores (yellow) by video number. Error bars represent 95% confidence intervals. Created with the ggpubr and ggplot package.

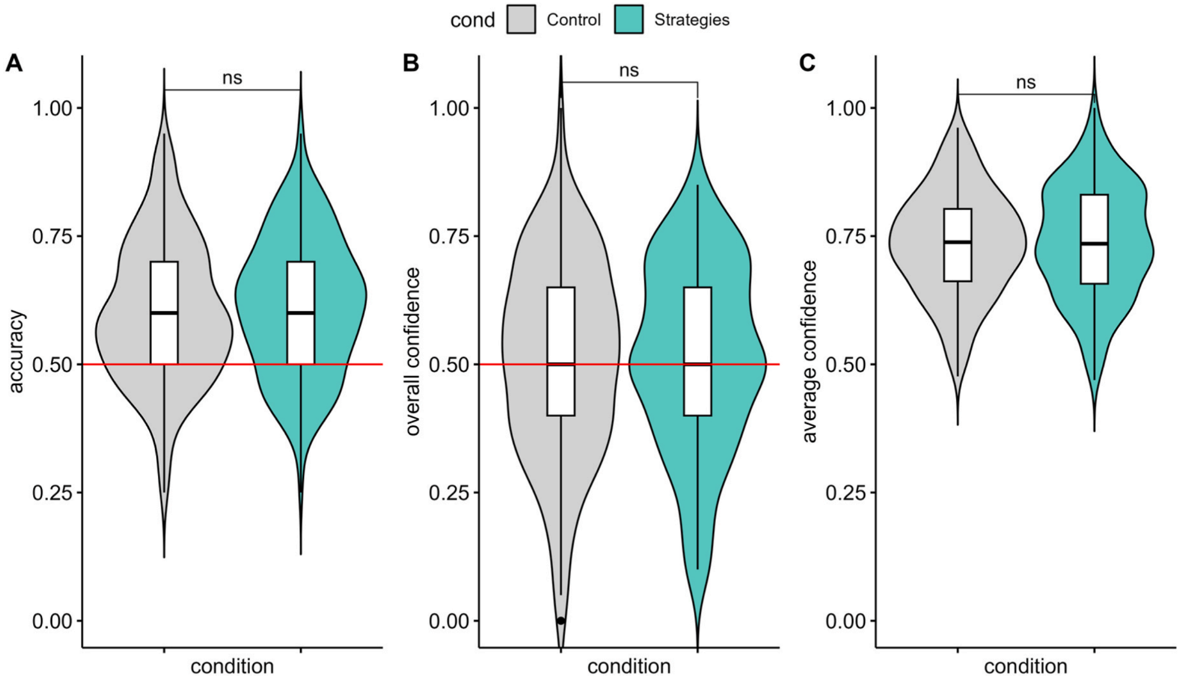

Fig. 3. Distribution of Accuracy and Confidence Scores by Treatment Group

Note. Violin plots of the distribution of (A) accuracy, (B) overall confidence and (C) average confidence per video by treatment group (x axis). Each violin plot shows medians and interquartile ranges. Violin plots show the data distribution and its probability density. Therefore, the wider areas in the plot indicate where the data is more common, while narrower areas indicate fewer common values. Created with the ggpubr and ggplot package.

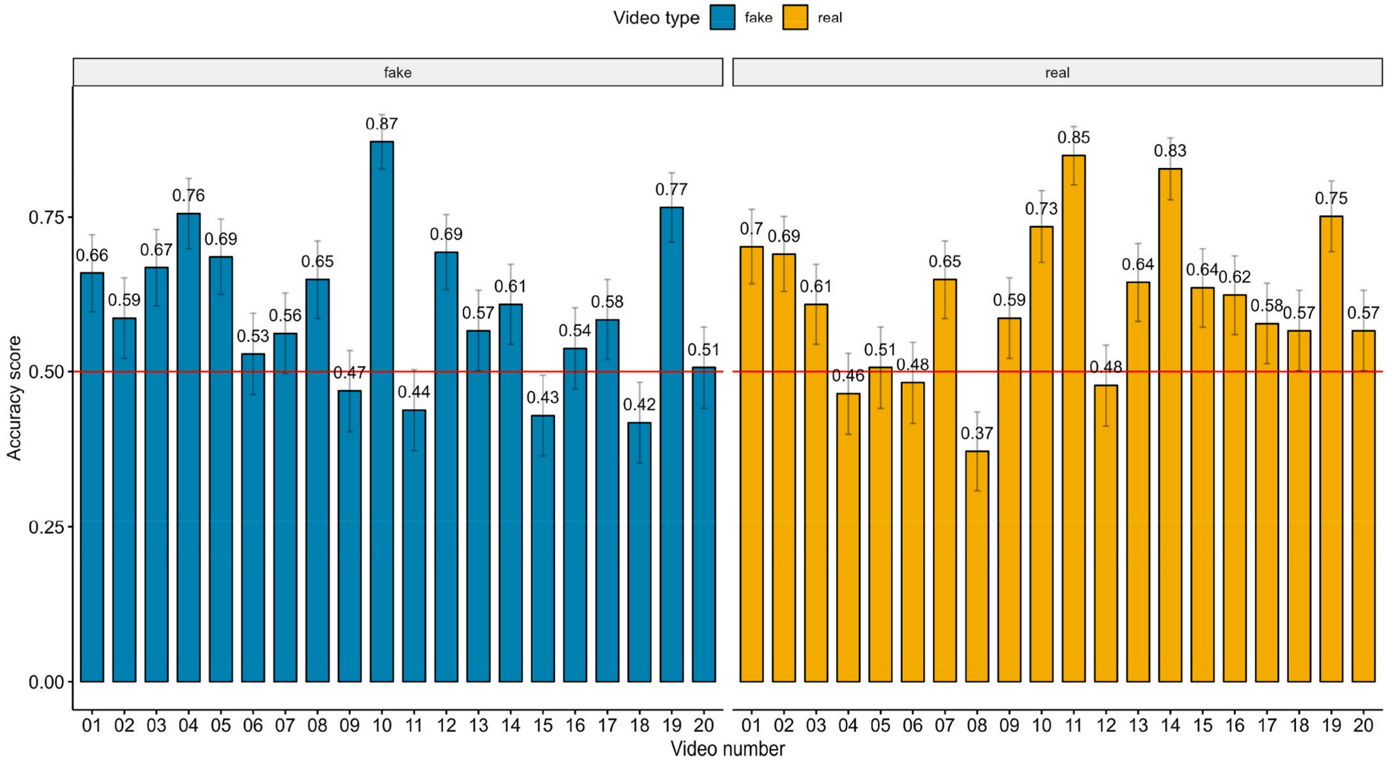

Fig. 4. Accuracy by Video Type

Note. The numbers presented in the plots are the average accuracy scores for fake (blue) and real videos (yellow) by video number. Error bars represent 95% confidence intervals. Created with the ggpubr and ggplot package.

participants were not given a monetary reward in exchange for participation. It is possible that participants in our study were already inherently motivated to correctly identify the videos as deepfake or real (as they were told that they will receive a score reflecting their performance), and therefore, were less likely to rely on heuristics when categorizing videos.

Our participants also performed better than those in the Kobis et al.'s (2021) study. Interestingly, while Kobis et al.'s (2021) intervention (aimed at increasing participants' motivation) did not improve performance on the detection task, overall level of motivation still positively predicted participants' performance (as they report in the supplementary material). Differences in performance between recruited and non-recruited participants were also observed in Groh et al.'s (2021) study. Participants who were recruited from Prolific accurately identified 66% of the videos, while non-Prolific participants (those who stumbled onto the study website on their own) performed better—accurately identifying 69% of videos. Groh et al. (2023) does not report if participants were compensated or not, although Prolific participants are usually given a monetary reward. It is possible that differences in performance between our study and Kobis et al.'s (2021) could be due to the differences in participants' intrinsic motivation. Due to our recruitment approach, we believe it is likely that participants in the current study were highly motivated and performed to the best of their abilities in the detection activity.

4.2. Efficacy of the detection intervention

At first glance, deepfakes and other computer-generated images can look quite realistic. However, computer algorithms can produce artifacts (e.g., unusual shadows, incongruence between skin, hair and eyes). It has been suggested that enhancing public awareness of common deepfake artifacts is one avenue via which to improve deepfake detection. Conversely, our study found that informing participants of these artifacts did not improve detection accuracy. Similarities in performance between the control and intervention group could be partly explained by humans' specialized ability to visually process faces, regardless of "training". In other words, it is possible that, when motivated, people already engage in such processing. Several studies conducted by Groh and colleagues (Groh et al., 2021, 2022) provide evidence for this. For instance, when actors in deepfake videos are inverted, misaligned, or occluded, participants' ability to detect deepfakes greatly diminishes. It is also possible that artifact interventions are effective, but only when reinforced with a training exercise of some kind (e.g., in which participants are instructed to look for artifacts and then provided with immediate feedback on their success in identifying these artifacts). While the experimental invention was not found to increase detection performance, it did not appear to bolster participants' confidence in their detection ability (which would be problematic for an ineffective intervention).

4.3. Participant confidence in detecting deepfakes

In terms of confidence to detect deepfakes (regardless of experimental group), Kobis et al. (2021) found that participants overestimated their abilities to detect deepfakes and that poorer performers showed greater overconfidence in their detection abilities. In the current study, participants also overestimated their detection abilities at the individual video level. However, participants somewhat underestimated their detection abilities when asked about their overall performance (on average, participants guessed that their overall detection abilities were very close to chance). This could be indicative of a cognitive bias in which people have a feeling of certainty regarding what is directly in front of them, even if they believe their long-term detection ability to be poor. For example, it is possible that people are more inclined to rely on feeling when making judgements at the individual video level, but when asked to reflect on their overall performance, individuals are more likely to draw on base-rate information (participants in the study were aware that they would be likely to correctly categorize 10 out of 20 videos by chance alone). However, it should also be noted that per video confidence was assessed via a sliding scale (anchored by *50 = As confident as flipping a coin* and *100 = 100% sure*) whereas overall confidence was assessed via an open response format (in which participants had to enter the number of videos they think they guessed correctly out of 20). Thus, the difference between overall confidence and per video confidence may also be attributable to middle-response style bias—the tendency for participants to choose middle response categories on rating scales (i.e., avoiding responding on the extreme ends of a scale, Harzing, 2006)—on the per video confidence questions.

No evidence was found for a Dunning-Kruger effect in which poor performers are more inclined to believe themselves to be skilled detectors. In fact, small positive correlations were observed between detection accuracy and both measures of detection confidence. Further, plotting performance quartile against performance separately for actual and perceived performance (Fig. S2) did not show a pattern of results consistent with the Dunning-Kruger effect (in which we would expect the slope for actual performance to be greater than the slope for perceived performance; Gignac & Zajenkowski, 2020), nor was a U-shaped distribution observed when plotting confidence against accuracy (Fig. S1).

4.4. Limitations

A number of limitations should be considered when evaluating the findings of the current study. First, the method of recruitment and online study format somewhat inhibited experimental control. For example, the study was advertised on social media platforms where comments are open to the public. Some participants shared detection strategies in these comments sections. It is possible that potential participants read these tips prior to doing the experiment (thereby potentially reducing the efficacy of the intervention).

Second, there are sociodemographic and personal history variables which could influence detection ability (or even moderate the effect of experimental intervention) which were not measured as part of the current study. These include prior experience with deepfakes or detection-type tasks, vision or face processing impairments, and even race or ethnicity. For example, there is an "other-race bias" for facial identification, whereby people show greater accuracy in recognizing the faces of members of their own race compared to the faces of members of other racial groups (Lee & Penrod, 2022; Meissner & Brigham, 2001). Whether a similar bias exists for the detection of deepfakes is unknown.

Third, the artificial nature of the experimental situation should be acknowledged. In a real-world setting it is extremely unlikely that media consumers would be told that the video they are watching has a 50% chance of being a deepfake. Given that the experimental situation (by necessity) sensitized participants to the possibility that they were viewing a deepfake, the detection accuracy levels observed in the current study might be considered best-case estimates. Alternatively, it could be argued that in a real-world situation media consumers would have contextual information to draw on (e.g., the plausibility of the message content, the trustworthiness of the video source), in addition to visual artifacts, thereby making it easier to determine the authenticity of a video.

4.5. Future studies

Future studies should investigate alternative interventions to enhance deepfake detection. For example, a video presentation on artifact detection with examples may be a more engaging detection intervention than simply providing written detection tips. Above, we also suggest including a training element in which participants are given immediate feedback after a set of practice videos before doing a detection activity. Researchers should also consider ways in which they can modify future deepfake studies to increase their ecological validity (for example, by embedding videos among other content, as would be typical of a website) to determine detection accuracy under more realistic conditions. Finally, replication attempts of the current study should a) make attempts to avoid participants sharing detection tips in comments sections (if recruiting via social media), b) bring the measurement of per video confidence and overall confidence into alignment (to identify if differences between these variables simply reflect response biases), and c) consider investigating additional individual difference variables, such as race and prior exposure to deepfakes.

5. Conclusion

Our study demonstrates that the public's ability to detect deepfakes is generally poor (although above chance levels), even in the idealized situation in which individuals are explicitly informed that they will be presented with deepfakes. The findings cast doubt on whether simply providing the public with strategies for detecting deepfakes based on the observation of visual artifacts can meaningfully improve detection, given the lack of an effect for the experimental intervention. Worryingly, it appears that individuals may be overly optimistic regarding their abilities to ascertain the authenticity of individual videos. However, individuals appear to have a more realistic understanding of their detection abilities in the long run.

Credit author statement

Dr. Klaire Somoray: Conceptualization, Methodology, validity, Formal analysis, Investigation, Data curation, Writing – original draft, review and editing, Visualization, Project administration. Dr. Dan J Miller: Conceptualization, Methodology, validity, Formal analysis, Investigation, Data curation, Writing – original draft, review and editing, Visualization, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be available in our OSF Repository.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chb.2023.107917.

References

eSafety. (2022). *Deepfake trends and challenges: Position statement*. Australian Government. https://www.esafety.gov.au/industry/tech-trends-and-challenges/deepfakes.

Ackerman, P. L., Beier, M. E., & Bowen, K. R. (2002). What we really know about our abilities and our knowledge. *Personality and Individual Differences, 33*(4), 587–605. https://doi.org/10.1016/S0191-8869(01)00174-X

Bregler, C., Covell, M., & Slaney, M. (1997). Video Rewrite: Driving visual speech with audio. *Proceedings of the 24th annual Conference on Computer Graphics and Interactive Techniques*, 353–360. https://doi.org/10.1145/258734.258880

Cole, S. (2018). We are truly fucked: Everyone is making AI-generated fake porn now. *Vice*. https://web.archive.org/web/20190907194524/.

Doctorow, C. (2018). February 7). Reddit shuts down deepfakes subreddit, home to faceswapped pornography (and some other stuff). *Boing Boing*. https://boingboing.net/2018/02/07/case-closed-then.html?fbclid=IwAR0yDv3dgI6QPSEX2x53Ius281rw_4iBsmDL2fD9K8jN7puFzUllqoIyrfg.

Dunning, D. (2011). The Dunning–Kruger effect: On being ignorant of one's own ignorance. In J. M. Olson, & M. P. Zanna (Eds.), *Advances in experimental social psychology* (Vol. 44, pp. 247–296). Academic Press. https://doi.org/10.1016/B978-0-12-385522-0.00005-6.

Gignac, G. E., & Zajenkowski, M. (2020). The Dunning-Kruger effect is (mostly) a statistical artefact: Valid approaches to testing the hypothesis with individual differences data. *Intelligence, 80*, Article 101449. https://doi.org/10.1016/j.intell.2020.101449

Groh, M., Epstein, Z., Firestone, C., & Picard, R. (2022). Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences, 119*(1), Article e2110013119. https://doi.org/10.1073/pnas.2110013119

Groh, M., Epstein, Z., Picard, R., & Firestone, C. (2021). Human detection of deepfakes: A role for holistic face processing. *Journal of Vision, 21*(9), 2390. https://doi.org/10.1167/jov.21.9.2390

Habgood-Coote, J. (2023). Deepfakes and the epistemic apocalypse. *Synthese, 201*(3), 103. https://doi.org/10.1007/s11229-023-04097-3

Harzing, A.-W. (2006). Response styles in cross-national survey research: A 26-country study. *International Journal of Cross Cultural Management, 6*(2), 243–266. https://doi.org/10.1177/1470595806066332

Hoaglin, D. C., & Iglewicz, B. (1987). Fine-tuning some resistant rules for outlier labeling. *Journal of the American Statistical Association, 82*(400), 1147–1149. https://doi.org/10.1080/01621459.1987.10478551

Kobis, N. C., Dolezalova, B., & Soraperra, I. (2021). Fooled twice: People cannot detect deepfakes but think they can. *iScience, 24*(11), Article 103364. https://doi.org/10.1016/j.isci.2021.103364

Korshunov, P., & Marcel, S. (2020). *Deepfake detection: Humans vs. machines. arXiv preprint arXiv:2009.03155* https://arxiv.org/pdf/2009.03155.pdf.

Kruger, J., & Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology, 77*(6), 1121. https://doi.org/10.1037/0022-3514.77.6.1121

Lee, J., & Penrod, S. D. (2022). Three-level meta-analysis of the other-race bias in facial identification. *Applied Cognitive Psychology, 36*(5), 1106–1130. https://doi.org/10.1002/acp.3997

Meissner, C. A., & Brigham, J. C. (2001). Thirty years of investigating the own-race bias in memory for faces: A meta-analytic review. *Psychology, Public Policy, and Law, 7*(1), 3–35. https://doi.org/10.1037/1076-8971.7.1.3

Mirsky, Y., & Lee, W. (2021). The creation and detection of deepfakes: A survey. *ACM Computing Surveys, 54*(1). https://doi.org/10.1145/3425780. Article 7.

MIT Media Lab. (n.d.). Detect DeepFakes: How to counteract misinformation created by AI. https://www.media.mit.edu/projects/detect-fakes/overview/.

Nuhfer, E., Fleisher, S., Cogan, C., Wirth, K., & Gaze, E. (2017). How random noise and a graphical convention subverted behavioral scientists' explanations of self-assessment data: Numeracy underlies better alternatives. *Numeracy: Advancing Education in Quantitative Literacy, 10*(1). https://doi.org/10.5038/1936-4660.10.1.4. Article 4.

Passos, L. A., Jodas, D., da Costa, K. A., Júnior, L. A. S., Colombo, D., & Papa, J. P. (2022). *A review of deep learning-based approaches for deepfake content detection. arXiv preprint arXiv:2202.06095* https://arxiv.org/pdf/2202.06095.pdf.

Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). Faceforensics++: Learning to detect manipulated facial images. *Proceedings of the IEEE/CVF International Conference on Computer Vision*. https://openaccess.thecvf.com/content_ICCV_2019/papers/Rossler_FaceForensics_Learning_to_Detect_Manipulated_Facial_Images_ICCV_2019_paper.pdf.

Tolosana, R., Vera-Rodriguez, R., Fierrez, J., Morales, A., & Ortega-Garcia, J. (2020). Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion, 64*, 131–148. https://doi.org/10.1016/j.inffus.2020.06.014

Wakefield, J. (2022). *Deepfake presidents used in Russia-Ukraine war*. BBC. https://www.bbc.com/news/technology-60780142.

Westerlund, M. (2019). The emergence of deepfake technology: A review. *Technology Innovation Management Review, 9*(11), 39–52. https://www.timreview.ca/article/1282.