Computers in Human Behavior Reports 16 (2024) 100538

Contents lists available at ScienceDirect

Computers in Human Behavior Reports

journal homepage: www.sciencedirect.com/journal/computers-in-human-behavior-reports

Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers

Alexander Diel [a,b,*], Tania Lalgi [a,b], Isabel Carolin Schröter [a,b], Karl F. MacDorman [c], Martin Teufel [a,b], Alexander Bäuerle [a,b]

[a] Clinic for Psychosomatic Medicine and Psychotherapy, LVR-University Hospital Essen, University of Duisburg-Essen, Essen, Germany
[b] Center for Translational Neuro- and Behavioral Sciences, University of Duisburg-Essen, Essen, Germany
[c] Luddy School of Informatics, Computing and Engineering, Indiana University, Indianapolis, USA

ARTICLE INFO

Keywords:
Accuracy
AI-generated content
Deepfake
Human detection
Synthetic face

ABSTRACT

Deepfakes are AI-generated media designed to look real, often with the intent to deceive. Deepfakes threaten public and personal safety by facilitating disinformation, propaganda, and identity theft. Though research has been conducted on human performance in deepfake detection, the results have not yet been synthesized. This systematic review and meta-analysis investigates human deepfake detection accuracy. Searches in PubMed, ScienceGov, JSTOR, Google Scholar, and paper references, conducted in June and October 2024, identified empirical studies measuring human detection of high-quality deepfakes. After pooling accuracy, odds-ratio, and sensitivity (*d'*) effect sizes ($k = 137$ effects) from 56 papers involving 86,155 participants, we analyzed 1) overall deepfake detection performance, 2) performance across stimulus types (audio, image, text, and video), and 3) the effects of detection-improvement strategies. Overall deepfake detection rates (*sensitivity*) were not significantly above chance because 95% confidence intervals crossed 50%. Total deepfake detection accuracy was 55.54% (95% CI [48.87, 62.10], $k = 67$). For audio, accuracy was 62.08% [38.23, 83.18], $k = 8$; for images, 53.16% [42.12, 64.64], $k = 18$; for text, 52.00% [37.42, 65.88], $k = 15$; and for video, 57.31% [47.80, 66.57], $k = 26$. Odds ratios were 0.64 [0.52, 0.79], $k = 62$, indicating 39% detection accuracy, below chance (audio 45%, image 35%, text 40%, video 40%). Moreover, *d'* values show no significant difference from chance. However, strategies like feedback training, AI support, and deepfake caricaturization improved detection performance above chance levels (65.14% [55.21, 74.46], $k = 15$), especially for video stimuli.

1. Introduction

1.1. Deepfakes

Advances in artificial intelligence (AI) have enabled the development of algorithms that produce content that is indistinguishable from the real world. Specifically, generative adversarial networks (GAN), trained on real-world data, can synthesize digital content using deep learning algorithms. GAN-generated content used to deceive the viewer into believing that it is real is called *deepfake*, a portmanteau of the words *deep learning* and *fake* (Chadha, Kumar, Kashyap, & Gupta, 2021; Lyu, 2020; Rana, Nobi, Murali, & Sung, 2022; Seow, Lim, Phan, & Liu, 2022). The term was first used in 2017 when a user of the social news and discussion website *Reddit* synthesized pornographic content of celebrities (Chadha et al., 2021; Lyu, 2020). Its misuse has since become popular and spread into other areas, such as the creation of fake political videos, nonconsensual pornography, fraud, or disinformation (Farid, 2022; Seow et al., 2022). Techniques employed to create deepfakes include face–body swaps, voice swaps, text-to-speech (TTS) to replace voices, face morphing, lip-syncing, and text generation (Chadha et al., 2021; Farid, 2022; Lyu, 2020).

Deepfakes attracted increased public attention in 2018 when a deepfake video began circulating of former U.S. president Barack Obama expressing controversial views (Homeland Security, 2022; Seow et al., 2022; Vaccari & Chadwick, 2020). Deepfakes have since been repeatedly misused in politics, for example, to shift public opinion before elections (BSI, 2024; Homeland Security, 2022; Whyte, 2020). Synthetic voices have also been used for financial fraud or identity theft (Bateman, 2020; Strupp, 2019). In the U.S., losses from AI fraud are estimated to exceed $12.5 billion in 2023 (Katanich, 2024). Deepfake

* Corresponding author. Clinic for Psychosomatic Medicine and Psychotherapy, LVR-University Hospital Essen, University of Duisburg-Essen, Essen, Germany.
E-mail address: alexander.diel@lvr.de (A. Diel).

https://doi.org/10.1016/j.chbr.2024.100538
Received 6 August 2024; Received in revised form 21 November 2024; Accepted 22 November 2024
Available online 30 November 2024

technology has been misused to create pornographic content containing a fake version of a real, often famous, person. Around 4000 celebrities have fallen victim to AI-generated pornography (Panda Security, 2024). An analysis by Sensity AI in 2021 has found that 90–95% of deepfake content consists of nonconsensual pornography (Hao, 2021). Ray reported that in 2020, one specific AI bot had created pornographic content containing over 100,000 women (Ray, 2020). Deepfake pornography has also been used to target children and adolescents (Winnard, 2024). Furthermore, AI-based text generators have been used to generate sexual stories involving children (Simonite, 2021).

Forged content is a long-standing security threat (Piva, 2013; Rocha, Scheirer, Boult, & Goldenstein, 2011). Examples of fraudulent content include synthesized faces on fake passports (Robertson et al., 2018), manipulated photos in journalism (Hadland, Cambell, & Lambert, 2015), misinformation in news (Adams, Osman, Bechlivanidis, & Meder, 2023), and false evidence presented in court (Amerini et al., 2013). Evidently, humans have trouble differentiating real content from forgeries (Nightingale, Wade, & Watson, 2017; Sanders, Ueda, Yoshikawa, & Jenkins, 2019; Schetinger, Oliveira, da Silva, & Carvalho, 2017). With recent technological advancements in the form of deepfakes, synthesized content resembles reality ever more closely. Given the ease of creating deepfakes and disseminating them through social media, believable forged content can spread quickly, causing great harm.

1.2. Types of deepfake content

Deepfake content can be categorized based on four modalities: audio, image, text, and video (Farid, 2022; Khanjani, Watson, & Janeja, 2023). Readily available *deepfake audios* created with AI-based TTS systems have become increasingly humanlike (Diel & Lewis, 2024; Müller, Pizzi, & Williams, 2022; Zhou, Ling, & King, 2020). Deepfakes find various uses, for example, in entertainment or marketing (Farid, 2022). Despite constructive applications like synthesizing the voice of a person who has lost theirs (i.e., voice banking; Judge & Hayton, 2022), AI-generated voices can be misused to impersonate, commit fraud or scientific misconduct, or spread misinformation (Farid, 2022; Khanjani et al., 2023; Mai, Bray, Davies, & Griffin, 2023; Strupp, 2019; Suwajanakorn, Seitz, & Kemelmacher-Shlizerman, 2017). For example, AI impersonation of a politician in political propaganda is not reliably detected by humans (Groh, Epstein, Picard, & Firestone, 2021).

Deepfake images can appear in fake photographic evidence and false identities (Caldwell, Andrews, Tanay, & Griffin, 2020). Research has focused mainly on the detection of synthesized human faces (Boyd, Tinsley, Bowyer, & Czaijka, 2023; Bray, Johnson, & Kleinberg, 2023; Cooke et al., 2024; Holmes, Banks, & Farid, 2016; Hulzebosch, Ibrahimi, & Worring, 2020; Lu et al., 2024; Mader, Banks, & Farid, 2017; Nightingale & Farid, 2022; Rössler et al., 2019; Shen, Richard Webster, O'Toole, Bowyer, & Scheirer, 2021; Tucciarelli, Vehar, Chandaria, & Tsakiris, 2022a). Deepfakes of human faces can be deceptive, appearing more trustworthy and authentic than their real counterparts (Nightingale & Farid, 2022; Tucciarelli et al., 2022a). Disruption of specialized perceptual processing for human faces hampers the ability to detect deepfakes, indicating that specialization supports detecting deepfakes (Groh, Epstein, Picard, & Firestone, 2021). In this vein, several studies trained human participants with feedback to improve their deepfake detection accuracy (Holmes et al., 2016; Mader, Banks, & Farid, 2017; Nightingale & Farid, 2022).

Deepfake texts are synthesized texts usually generated by a large language model (LLM). AI-generated text poses a threat in the form of fake news articles (Aïmeur, Amri, & Brassard, 2023; Hamed, Ab Aziz, & Yaakub, 2023; Keya, Shajeeb, Rahman, & Mridha, 2023; Twomey et al., 2023), comments, tweets, and reviews (Fagni, Falchi, Gambini, Martella, & Tesconi, 2021; Rupapara et al., 2021; Weiss, 2019). LLMs can be used to create misinformation that manipulates public opinion. Furthermore, fake academic texts created by university students or academics cannot be reliably detected as such, highlighting the threat of AI-generated text to academic integrity (Elali & Rachid, 2023; Gao et al., 2023; Hakam et al., 2024; Ibrahim et al., 2023; Májovský, Černý, Kasal, Komarc, & Netuka, 2023; Popkov & Barrett, 2024; Odri & Yoon, 2023; Rashidi, Fennell, Albahra, Hu, & Gorbett, 2023).

Deepfake videos possess various uses and risks (Yu, Xia, Fei, & Lu, 2021). Deepfake videos can combine visual (e.g., human bodies and faces) and auditory (e.g., voices) AI-generated content. The detection of deepfake videos, such as political deepfakes, has been the focus of multiple studies (Cooke, Edwards, Barkoff, & Kelly, 2024; Doss et al., 2023; Groh, Epstein, Firestone, & Picard, 2022; Köbis, Doležalová, & Soraperra, 2021; Korshunov & Marcel, 2020; Mittal, Sinha, Swaminathan, Collomosse, & Manocha, 2023; Somoray & Miller, 2023).

1.3. Deepfake detection strategies

Several studies have tested strategies to improve our ability to detect deepfake content. Strategies include AI support for detection decisions (Groh et al., 2022), attentional strategies (Bray et al., 2023), feedback training (Holmes et al., 2016), financial incentives (Köbis et al., 2021), human collaboration (Uchendu et al., 2023), creating deepfake caricatures (Fosco et al., 2022), and raising awareness (Tucciaelli et al., 2022). However, a general synthesis of detection improvement strategies and a systematic comparison of these strategies are still lacking.

GAN-created deepfake content has found several constructive uses. For example, deepfake faces may be used as stimuli for research in face processing (Reiner et al., 2024), emotion processing (Becker et al., 2024), health informatics (Dai & MacDorman, 2021), and the social sciences (Eberl, Kühn, & Wolbring, 2022; Tucciarelli et al., 2022a). In addition, deepfake "twin doctors" may be used as embodied conversational agents in remote interactions with clinical patients (Zalake, 2023).

Several AI algorithms for detecting deepfakes have been devised and tested, summarized in multiple reviews and meta-analyses (Heidari, Navimipour, Dag, & Unal, 2024; Juefei-Xu et al., 2022; Rana et al., 2022; Stroebel, Llewellyn, Hartley, Ip, & Ahmed, 2023; Whittaker, Mulcahy, Letheren, Kietzmann, & Russell-Bennett, 2023; Zotov, Dremliuga, Borshevnikov, & Krivosheeva, 2020). Despite interest in *human* deepfake detection, the topic is underexplored (Lyu, 2020). Although AI algorithms can detect deepfakes, they are not readily available to the public. Thus, they do not protect people when they are confronted with deepfakes on the Internet—in false advertising, identity-theft pornography, political propaganda, misinformation, and fraudulent academic articles (Caldwell et al., 2020; Campbell, Plangger, Sands, & Kietzmann, 2021; Fink, 2019; Hamed et al., 2023; Ibrahim et al., 2023; Keya et al., 2023).

Several national information security agencies warn of the risk of deepfakes and suggest strategies for consumers to mitigate deception, such as raising awareness and being cautious with AI artifacts (Bundesministerium Inneres, 2024; BSI, 2024; Canadian Security Intelligence Service, 2023; Homeland Security, 2022). However, it is unclear whether these strategies are effective. In general, the research on *human* performance in deepfake detection has been sparse. The authors were unable to find any systematic reviews or meta-analyses on the topic. To fill this gap, this work synthesizes research findings on human deepfake detection performance.

1.4. Hypotheses

This work presents the first meta-analysis on human deepfake detection performance. It aims to synthesize the evidence on the human ability to detect deepfakes, both collectively and across deepfake audios, images, texts, and videos (Groh et al., 2024; Khanjani et al., 2023). In addition, the effects of strategies to enhance deepfake detection are reviewed. Hence, the following hypotheses are tested:

1. Humans can accurately detect deepfakes.

2. Humans can accurately detect deepfakes across four modalities: audio, image, text, and video.
3. Strategies to improve human deepfake detection accuracy increase human performance.

2. Methods

This meta-analysis has been conducted and is reported in accordance with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA; Page et al., 2021), adapted to the present study's designs and the journal's requirements. It was preregistered in May 2024 at https://doi.org/10.17605/OSF.IO/7W645. Datasets and analysis scripts are available at https://osf.io/hnf8g/.

2.1. Inclusion criteria

This work's goal is to synthesize research on human deepfake detection performance; hence, papers were included by the following criteria:

1) novel empirical research (excluding reviews, surveys, essays, and theoretical works),
2) data from human participants (excluding research only with AI detection algorithms),
3) high-quality deepfake stimuli (i.e., stimuli generated using deep learning algorithms, excluding stimuli described as unrealistic, low quality, noisy, or not created using a deep learning algorithm, such as manipulated stimuli),
4) detection performance measures (i.e., measures that directly assess the accuracy of detection, either using two-alternative tasks or transforming detection data into a binary result, excluding measures not directly indicating detection performance, such as Likert-scales on perceived naturalness or authenticity, because participants may judge a known deepfake to be authentic-looking), and
5) sufficient information to calculate effect sizes, including raw data. Authors of papers with insufficient information were contacted for more information. If the authors did not respond after one week, a reminder e-mail was sent, and after three weeks, the paper was excluded.

2.2. Search and selection

Searches were conducted in June 2024 and October 2024. The PubMed, ScienceGov, and Jstor databases were searched using these search terms (or adapted versions): ("human" OR "perception" OR "perceptual") AND ("deepfake" OR "deep-fake" OR "deep fake" OR "artificially generated" OR "AI-generated" OR "synthetic face") AND ("detection" OR "recognition"). In addition, reference sections of relevant published literature were screened for papers. Google Scholar was also used as a source. Due to Google Scholar's low specificity, only the first 500 results were screened. Three independent reviewers assessed and selected the papers according to the inclusion criteria. Out of 1,181 studies found, 56 were selected for the meta-analysis. Fig. 1 depicts the selection process.

2.3. Data extraction and analysis

Three independent reviewers performed data extraction. To increase the results' robustness according to the PRISMA guidelines (Page et al., 2021), the analysis was performed using different effects: proportions, odds ratios (ORs), and sensitivity indices (*d'*s) based on reported confusion matrix values (i.e., hits, false alarms, misses, and correct rejections). Given that some studies did not report all relevant data or provide access to the raw data (e.g., only total accuracy ratings instead of accuracy for real and deepfake stimuli separately), slightly different sets of studies were included for different analyses. Therefore, analyzing and reporting three different measures of accuracy (proportions, ORs, and *d'*s) will summarize the relevant literature while also increasing robustness to estimate sensitivity. The reported data includes the effect size, its 95% confidence interval in brackets, and the number of measures (*k*) from which it was derived.

Data analysis was conducted using MedCalc (ver. 20.217.0.0) and R

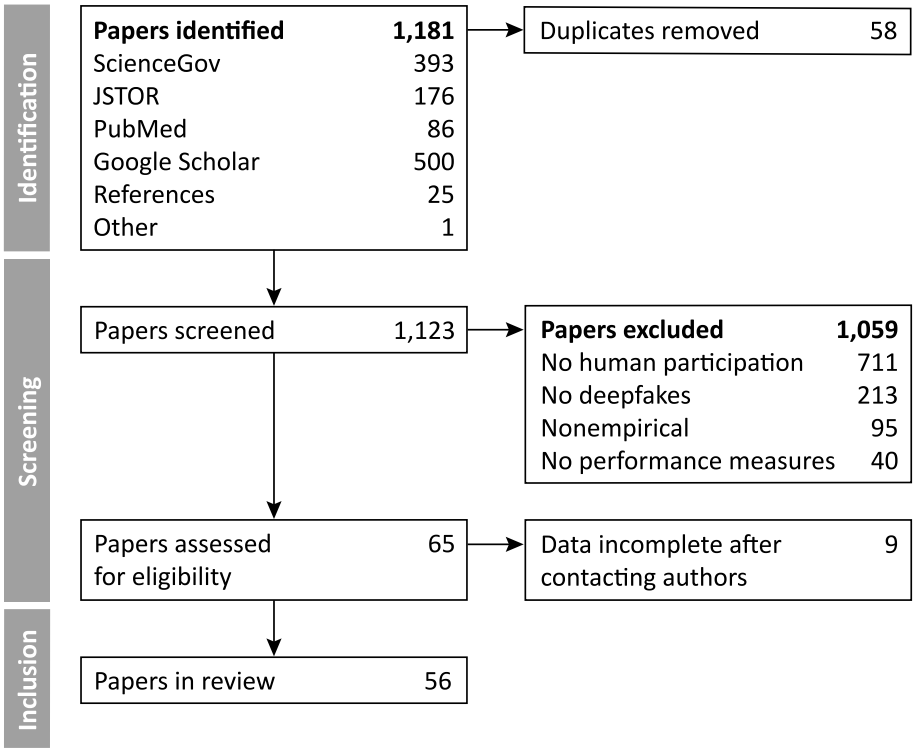

Fig. 1. Flowchart of the study selection process.

(ver. 4.1.2). The *forest* function from the *metafor* R package was used to depict the effect sizes (Viechtbauer, 2010). Heterogeneity analysis was performed by calculating Cochran's Q (χ^2) significance test and I^2 statistics (Higgins, Thompson, Deeks, & Altman, 2003). A significant Cochran's Q and I^2 values above 50% indicate heterogeneity (Higgins et al., 2003). Heterogeneity of the results was expected and mitigated by conducting random-effect models and subgroup analyses, decided a priori.

Meta-analysis models were random effects models with paper as the random effect and measure as the fixed effect. A random-effect model was chosen to control for paper-group effects for papers with multiple studies, as varied sample characteristics and stimulus modality, quality, and content were expected. The random effects model was calculated according to DerSimonian and Laird (1986), with studies weighted by sample size and number of stimuli. 95% confidence intervals were calculated as 1.96 times the standard error and were reported with the measures. Confidence intervals were used to interpret the significance of results. They indicate no significant difference from chance when they cross 50% for proportions, 1 for ORs, or 0 for d'. For each analysis, studies were weighted according to sample and stimulus size.

2.3.1. Proportions

Proportions of correct identifications are defined as the rate of correctly identifying a stimulus (i.e., the hit rates and correct rejections). Proportions were synthesized across studies for three variables: total accuracy (including across real and deepfake stimuli), accuracy for real stimuli, and accuracy for deepfake stimuli. Proportions were used as they are the easiest to understand and the most frequently reported. Variances were stabilized using the Freeman-Tukey arcsine square root transformation (Freeman & Tukey, 1950).

Although total proportions were calculated, which include both deepfake and real stimuli, these values can be misleading due to dataset imbalance and the aggregation of individual class contributions (He & Garcia, 2009; Tharwat, 2021). Therefore, results were interpreted using deepfake stimuli accuracies and their confidence intervals.

2.3.2. Odds ratios

OR values were calculated to assess the odds of correctly detecting deepfake stimuli relative to real stimuli as a measure of the human ability to detect deepfakes. ORs were calculated using this formula:

$$OR = \frac{ad}{bc} \tag{1}$$

where a is the number of deepfakes detected, b is the number of deepfakes missed, c is the number of real stimuli detected, and d is the number of real stimuli missed (Borenstein et al., 2009). The meta-analysis on ORs was calculated by the Mantel-Haenszel method (Mantel & Haenszel, 1959).

2.3.3. Sensitivity index d'

Proportions do not account for response errors or biases. For example, high accuracy in deepfake detection may result from a response bias towards labeling any stimulus as deepfake, which would not represent true discrimination ability. Signal detection theory uses the sensitivity index d' to control for response bias by calculating the separation between the deepfake (signal) and real stimulus (noise) distributions (Macmillan, 2002).

For each study, the sensitivity index d' was calculated:

$$d' = z(H) + z(1 - FA) \tag{2}$$

where $z(H)$ are the z-scores for the hit rate (correctly identified deepfake stimuli) and $z(1 - FA)$ are the z-scores of 1 minus the false alarm rate (1 minus real stimuli incorrectly identified as deepfake).

For each study, the variance of d' was estimated using the following formula:

$$Var(d') = \frac{1}{n_s H(1-H)} + \frac{1}{n_n FA(1-FA)} \tag{3}$$

where n_s and n_n are the number of signal and noise trials, respectively, and H and FA are hit and false alarm rates.

For the meta-analysis, study-level d' values were synthesized into combined d' values. Combined d' values and their variances were calculated using the formulas

$$Combined\ d' = \frac{\sum wd'}{\sum w} \tag{4}$$

and

$$Var(Combined\ d') = \frac{1}{\sum w} \tag{5}$$

where w is each study's weight calculated as the inverse of the variance of d':

$$w = \frac{1}{Var(d')} \tag{6}$$

2.3.4. Publication bias

Publication bias was assessed by funnel plot asymmetry and p-curve analysis for proportion and OR values. Funnel plot asymmetry (assessed by Egger's test) would show a pattern of smaller studies having larger effect sizes, which could occur due to a bias towards publishing significant rather than nonsignificant results (Borenstein, Hedges, Higgins, & Rothstein, 2009). P-curve analysis was used to investigate whether a bias existed to publish results with p-values just below the significance threshold (i.e., 0.05), indicating p-hacking.

3. Results

This meta-analysis includes 56 papers, listed in Table A1 of the Appendix. Figure A1 provides a confusion matrix of hit and miss rates of real and deepfake stimuli across all studies.

3.1. Heterogeneity

Heterogeneity analysis revealed significant heterogeneity for proportions, $Q(84) = 636{,}237.93$, $p < .001$, $I^2 = 99.99\%$ [99.99, 99.99], and ORs, $Q(61) = 116{,}847.43$, $p < .001$, $I^2 = 99.95$ [99.95, 99.95]. The random effects model on proportions had a fit of $R^2 = 0.57$. and on ORs of $R^2 = 0.47$.

3.2. Publication bias

Publication bias was assessed by funnel plot analysis on proportion and OR values (Fig. 2). Egger's tests indicated no significant asymmetry for proportions ($t(85) = 15{,}94$, $p = .161$) or ORs ($t(45) = 6.14$, $p = .254$). Hence, no evidence of a publication bias was found.

P-curve analysis revealed significant right-skewedness ($z = -65.93$, $p < .001$) and nonsignificant flatness ($z = 68.25$, $p = 1.00$). Hence, the p-curve analysis did not indicate publication bias. The p-curve is depicted in Fig. 3.

3.3. Proportions

Across all trials, 47.42% of stimuli were real, and 52.58% were deepfake. Across all studies and stimuli, participants performed better at detecting real stimuli, correctly detecting them 68.08% [64.74, 71.26] of the time, $k = 64$ effect sizes. However, they only correctly detected deepfake stimuli 55.54% [48.87, 62.10] of the time, $k = 67$. The overall mean accuracy was 60.60% [56.51, 64.62], $k = 85$. Random-effects analysis with paper grouped by stimulus type is shown in Fig. 4.

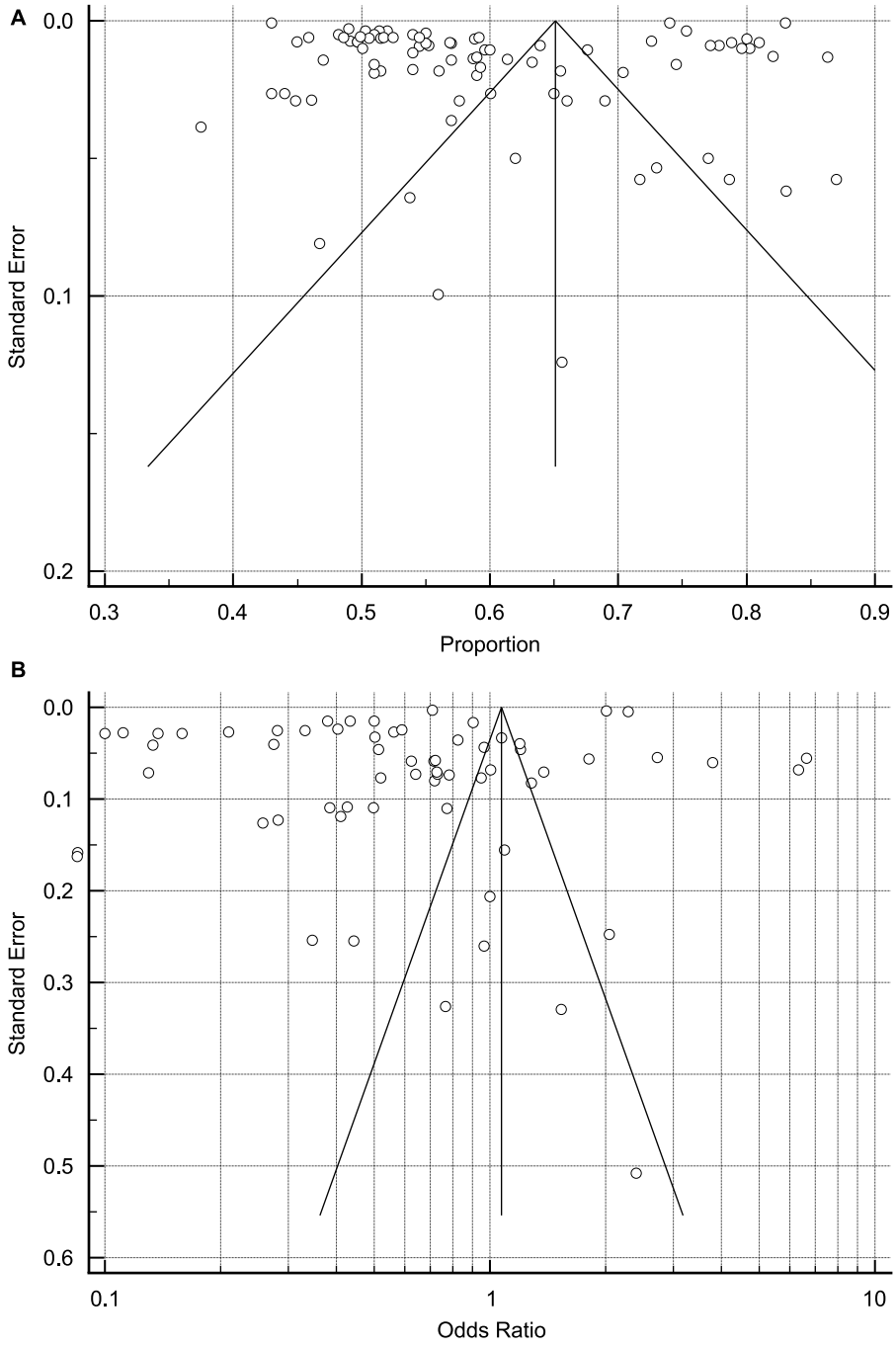

Fig. 2. Funnel plots depicting effects plotted against standard errors (reverse scaled). Fig. 2(a) depicts the funnel plot for proportions and Fig. 2(b) for OR values.

Overall, deepfake stimuli were not detectable.

3.3.1. Modality-level analysis

For audio, total accuracy was 63.11% [49.64, 75.61], $k = 9$. Accuracy was 70.67% [65.43, 75.65], $k = 8$, for real voices and 62.08% [38.23, 84.18], $k = 7$, for deepfake voices.

For images, total accuracy was 58.04% [53.31, 62.70], $k = 26$. Accuracy was 68.46% [64.39, 72.40], $k = 18$) for real images and 53.16% [42.12, 64.64], $k = 18$, for deepfake images.

For text, total accuracy was at 58.00% [50.68, 64.94], $k = 17$. Accuracy was 66.81% [64.04, 69.53], $k = 14$) for real text and 52.00% [37.42, 65.88], $k = 15$, for deepfake text.

For videos, total accuracy was 63.26% [57.73, 68.62], $k = 33$. Accuracy was 68.00% [60.12, 74.63], $k = 23$) for real videos and 57.31% [47.80, 66.57], $k = 26$, for deepfake videos.

3.3.2. Effects of strategy

Using a strategy to improve deepfake detection increased general accuracy to 63.31% [56.12, 69.31], $k = 17$. While accuracy did not show a noteworthy increase for real stimuli, 67.21% [64.00, 70.62], $k = 13$, accuracy for deepfake stimuli increased by about 10% with 65.14% [55.21, 74.46], $k = 15$. Hence, applying deepfake detection strategies improves the correct identification of deepfake stimuli. Results for all proportion analyses are summarized in Fig. 4.

3.4. Odds ratios

OR analysis on the odds of detecting a deepfake relative to the odds

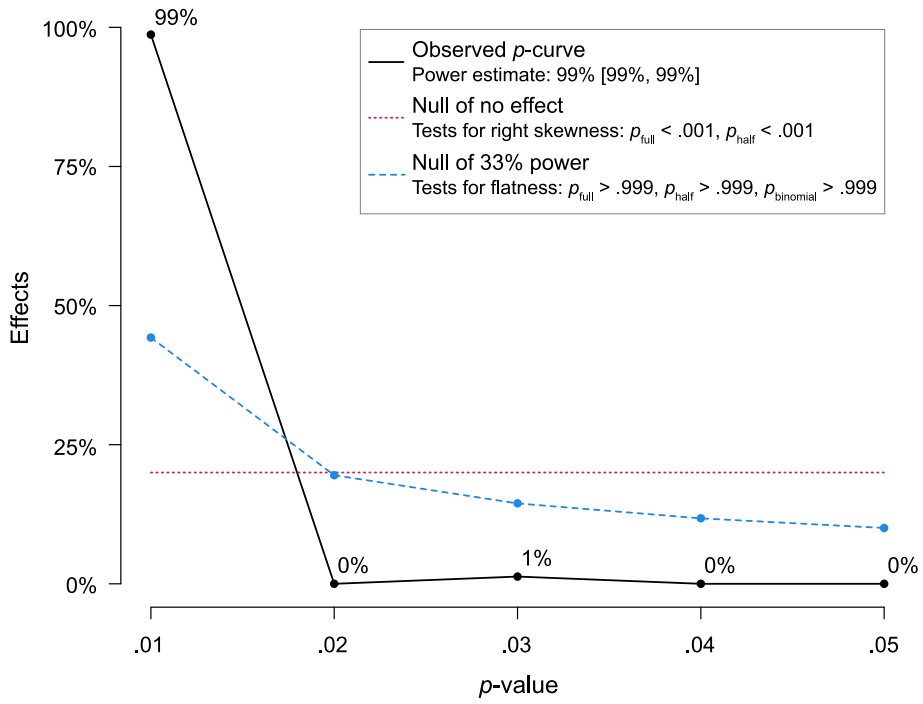

Fig. 3. *P*-curve analysis across 58 effects. The curve's shape does not indicate publication bias.

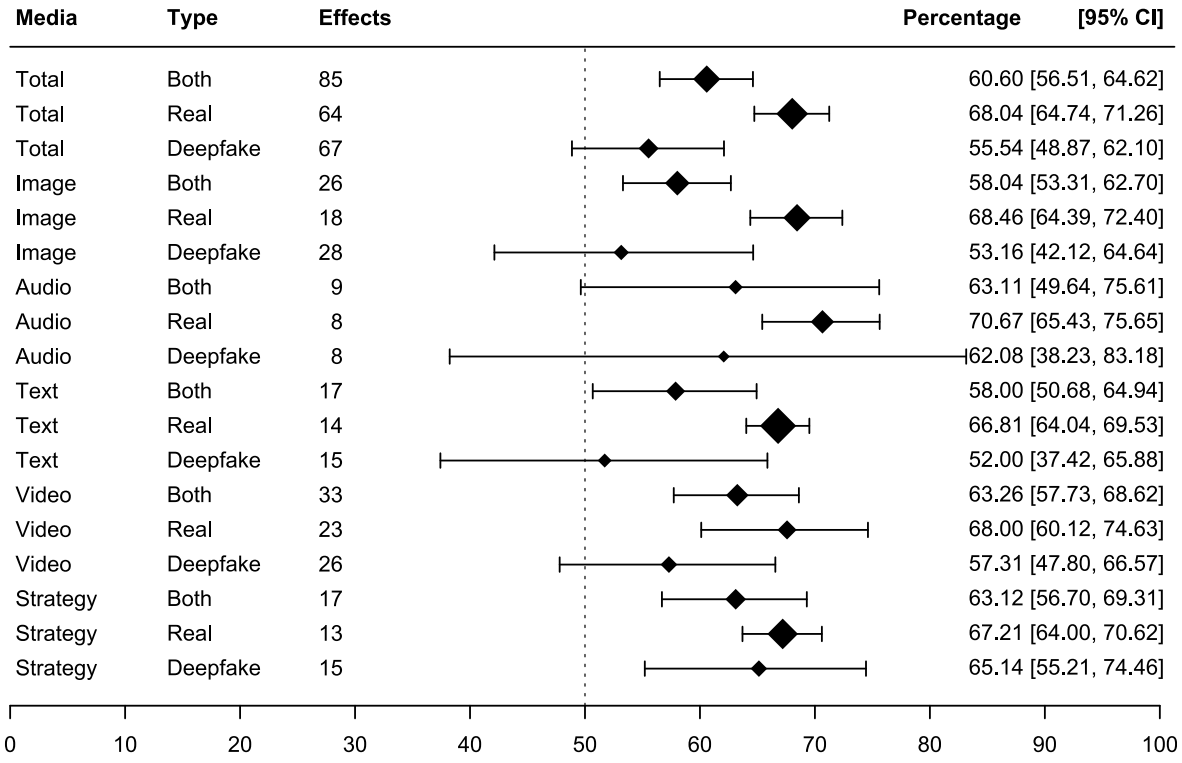

Fig. 4. The probability of correctly identifying deepfake and real stimuli, grouped by condition. Error bars indicate confidence intervals. The dashed line indicates chance (50%). Error bars indicate 95% confidence intervals.

of missing a deepfake revealed an OR at 0.64 [0.52, 0.79], $k = 62$, indicating that the odds of detecting a deepfake were 39% and the odds of missing a deepfake were 61%. Modality-level analysis revealed the odds of detecting deepfakes across media types. For audio, an OR of 0.81 [0.31, 2.09], $k = 8$, indicates 45% odds of detection versus 55% odds of missing. For images, an OR of 0.53 [0.34, 0.53], $k = 18$, indicates 35% odds of detection versus 65% odds of missing. For text, an OR of 0.66 [0.37, 1.15], $k = 15$, indicates 40% odds of detection versus 60% odds of missing. For videos, an OR of 0.68 [0.53, 0.86], $k = 21$, indicates 40% odds of detection versus 60% odds of missing.

Finally, the effects of strategy on OR showed an increased OR of 0.81 [0.66, 1.00], $k = 12$, indicating that when applying a strategy to improve deepfake detection, the odds of detecting a deepfake become closer to the odds of missing one. The results of the OR analyses are summarized

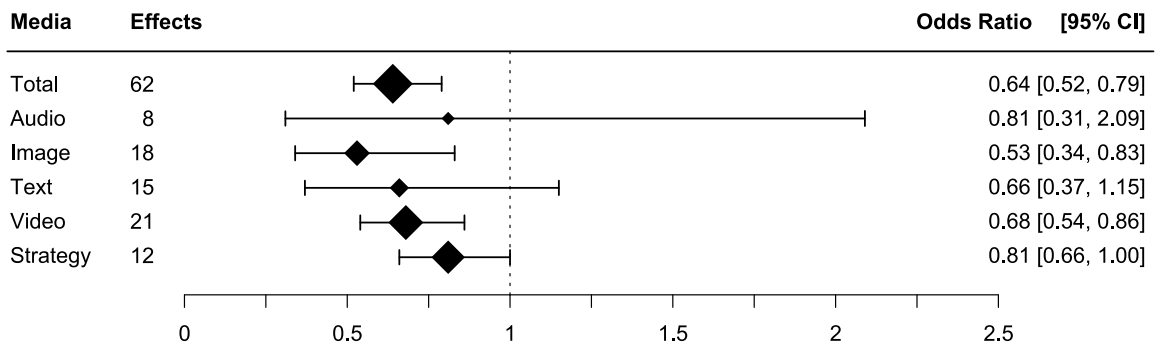

Fig. 5. The odds of detecting a deepfake relative to the odds of missing a deepfake (OR) by stimulus condition. Error bars indicate 95% confidence intervals.

in Fig. 5.

3.5. Analysis of *d'*

Finally, *d'* analysis revealed results like those for proportions and ORs. Combined *d'* was at 0.67 [–0.01, 1.36], $k = 68$, indicating a low sensitivity due to the confidence intervals overlapping with 0. The sensitivity index *d'* was 1.25 [–0.01, 2.50], $k = 8$, for audio, 0.54 [–0.01, 1.10], $k = 29$, for images, 0.022 [–0.05, 0.44], $k = 16$, for text, and 1.74 [–0.01, 3.49], $k = 26$, for videos.

When applying strategies to increase performance, *d'* increased slightly from the control of 0.17 [–0.02, 0.36], $k = 13$, to 0.22 [–0.02, 0.47], $k = 11$. Nevertheless, the combined *d'* remained close to chance. Sensitivity index results appear in Fig. 6.

3.6. Review of improvement strategies

Various strategies have been developed to improve the accuracy of deepfake detection. Due to the limited number of studies investigating specific strategies and several studies not reporting measures that could be synthesized (e.g., area under the curve or *d'* values without variance), a meta-analysis would not be sufficient to synthesize the results of the effects of strategy on performance. Hence, a review of the applied strategies has been conducted. A summary of strategies and their results is shown in Table 1.

Six out of 17 studies trained participants using feedback, resulting in near consistent improvement in deepfake detection (Diel et al., 2024; Holmes et al., 2016; Hulzebosch, Ibrahimi, & Worring, 2020; Mader, Banks, & Farid, 2017; Müller, Pizzi, & Williams, 2022; however, see Nightingale & Farid, 2022). Feedback training uses operant conditioning, specifically of reward (through positive feedback) and punishment (through negative feedback) to reinforce the detection of features specific to deepfake content (e.g., artifacts). While most studies used visual stimuli, Müller, Pizzi, and Williams (2022) also found that feedback training had a positive effect for audio stimuli, indicating that the benefit of feedback training could be independent of modality.

Three studies applied the raising-awareness strategy to deepfakes (Ahmed et al., 2021; Köbis et al., 2021; Tucciarelli et al., 2022a). Although all three found improvements in deepfake detection, the change was only significant in one study (Tucciarelli et al., 2022a) and either not significant (Köbis et al., 2021) or not tested for significance in the rest (Ahmed et al., 2021). In addition, the increased performance remained close to chance.

Three studies used advice (e.g., on attending to deepfake features) with mixed results (Bray et al., 2023; Somoray & Miller, 2023; Tahir et al., 2021). While Bray et al. (2023) found an increase in detection performance, performance did not increase for Somoray and Miller (2023). Tahir et al. (2021) employed a training program with explicit advice on detecting deepfakes (e.g., by focusing on typical artifacts). They found a noteworthy increase in total accuracy.

Two studies applied support-based strategies: Groh et al. (2022) provided participants with AI decisions on an image, and Uchendu et al. (2023) let participants collaborate in groups. Both support strategies improved detection performance.

Finally, two studies used a deepfake caricaturization method to exaggerate deepfake artifacts (Fosco et al., 2023; Josephs et al., 2023). Whereas real face caricatures presented more distinctive features, a deepfake caricatures presented more detectable artifacts. Both studies reported a substantial increase in deepfake detection.

In general, the review found that strategies improved deepfake detection. Although feedback training showed the most consistent positive results, advice, AI support, group collaboration, and exaggeration also improved deepfake detection. Furthermore, exaggerating deepfake artifacts through caricaturization shows promise in greatly improving deepfake detection.

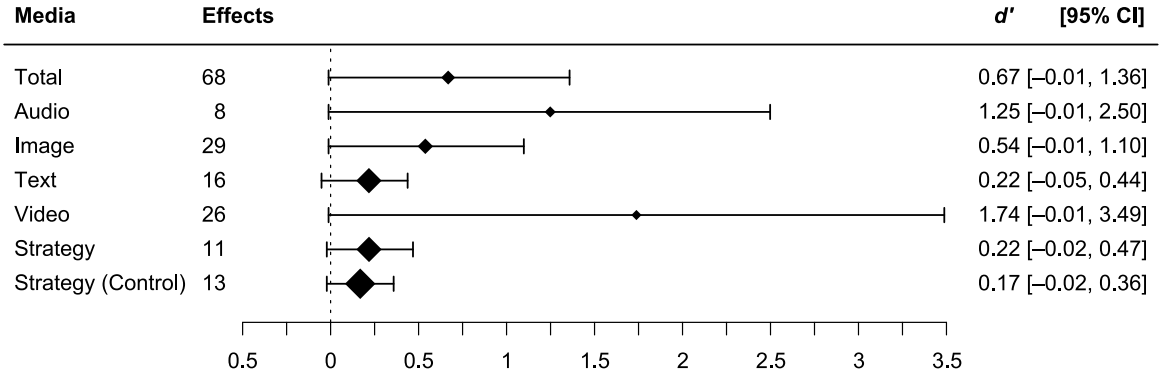

Fig. 6. Combined *d'* values across stimulus conditions. Error bars indicate 95% confidence intervals.

Table 1
Results of individual studies' strategies to improve deepfake detection. Results are divided into a control group (CG) and an intervention group (IG). AUC indicates the area under the curve.

Study	Strategy	Result type	Result
Groh et al. (2022)	AI support	AUC	CG: 0.936 IG: 0.982
Nightingale and Farid (2022)	Feedback training	*d'*	CG: −0.09 IG: 0.46
Mader, Banks, and Farid, 2017	Feedback training	*d'*	CG: 1.65 IG 1: 2.01 IG 2: 2.18
Hulzebosch, Ibrahimi, and Worring, 2020	Feedback training	Proportion (deepfake)	CG: 60.4% IG: 70.9%
Holmes et al. (2016)	Feedback training	*d'* on various resolution levels	CG: 1.45 to 1.68 IG: 1.42 to 1.91
Tucciarelli et al. (2022a	Deepfake awareness	*d'*	CG: 0.05 IG: 0.15
Köbis et al. (2021)	Deepfake awareness	Proportion; *d'*	CG: 57.6%; 0.44 IG: 67.4%; 0.51
Müller, Pizzi, and Williams, 2022	Feedback training	Proportion	IG 1: 67% IG 2 80%
Bray et al. (2023)	Advice: one time versus repeated	Proportion (deepfake)	CG: 51.75% IG 1: 62.25% IG 2: 69.10%
Boyd et al. (2023)	AI support	Proportion	CG: 56% IG: 61%
Uchendu et al. (2023)	Social support	Proportion (33% chance), incl. expert condition	CG nonexpert: 45% CG expert: 56% IG nonexpert: 51% IG expert: 69%
Somoray and Miller (2023)	Advice	Proportion	CG: 60% IG: 61%
Tahir et al. (2021)	Training with explicit advice	Proportion	CG: 58%–57% IG: 55%–88%
Ahmed, Miah, Bhowmik, and Sulaiman (2021)	Deepfake awareness	Proportion (deepfake)	Phase 1: 29% Phase 2 (post-intervention): 52%
Josephs, Fosco, and Oliva (2023)	Caricaturization	Proportions	For caricatures: 94.9%, 94.4%, 92.6%, 93.2%
Fosco et al. (2022)	Caricaturization	Proportions	Increase of accuracy by 14% (43% for a 5-s exposure)
Diel, Teufel, and Bäuerle (2024)	Feedback training	Proportions	Increase of accuracy from 45% to 65%; no change in control group

4. Discussion

Deepfake content is becoming increasingly widespread, realistic, and hard to detect. This meta-analysis, the first on human deepfake detection, found synthesized human performance to be at chance. Across all stimulus types (audio, image, text, and video), 95% confidence intervals crossed chance levels. Thus, deepfake detection performance is not significantly above chance for any modality or overall. Performance remains relatively consistent across stimulus types (audio, image, text, and video), with lower accuracy for detecting deepfakes than real stimuli. Finally, strategies to improve human deepfake detection generally succeed, though not consistently.

The funnel plot and *p*-curve analyses showed no publication bias. Publication bias arises when significant results are valued over nonsignificant ones, leading to their being disregarded (Sutton, 2009). However, a publication bias was not expected for human deepfake detection performance because of the societal impact of nonsignificant findings. The finding that humans detect deepfakes at chance levels underscores the threat of deception and manipulation deepfakes pose.

Meta-analyses of proportions, ORs, and *d'* values and the systematic review of strategies yielded similar results. The meta-analysis of correct identification proportions showed that detection performance tends to be higher for real stimuli than deepfakes. Furthermore, OR analysis showed that the odds of identifying a deepfake are about half the odds of missing a deepfake. Finally, *d'* analysis generally revealed low sensitivity in detecting deepfakes, with confidence intervals crossing chance levels. Proportion analysis further showed that applying strategies usually increases deepfake detection. This improvement is reflected in the ORs, which revealed that after applying a strategy, the odds of correctly identifying or missing a deepfake are almost the same. The review of strategies, which includes studies not included in the meta-analyses due to limited data availability, further showed an increase in accuracy after training.

Among studies with strategies, most used a training task with immediate feedback after participants decided whether a stimulus was real or fake. Combining different strategies may lead to further, incremental improvements. For example, feedback training may enhance perceptual strategies for detecting deepfakes, and training efficacy could be improved by providing users with additional advice on how to detect deepfakes. Performance can be further enhanced by letting trained users use AI. Tahir et al. (2021) found extensive training with explicit advice increased accuracy 30%. Amplifying artifacts by creating deepfake caricatures increased accuracy above 90%.

Performance for audio stimuli was higher in proportions and ORs than other metrics. The 95% confidence intervals crossed random chance levels and remained wide. Audio accuracy ranged from low (28%, Frank et al., 2023) to high (87%, Groh et al., 2024). This variation reflects heterogeneity in study factors influencing performance, such as deepfake stimulus quality and familiarity (e.g., random voices versus politicians' speeches). Thus, while the results do not support above-average deepfake detection for audio, they may depend on the kind of stimulus.

The large-scale experiments of Groh et al. (2022, 2024) reported higher detection accuracies than this meta-analysis. Groh et al. (2022) used deepfake videos, which are multimodal, incorporating images, movement, and voice. As AI may generate noticeable errors in any of these modalities, the chance of detecting errors in video is higher than in single-modality stimuli. Moreover, Groh et al. (2024) derived their audio, text, and video stimuli from videos of Donald Trump and Joseph Biden. The higher accuracy of their results may be due to hightened sensitivity to to distortions in familiar faces than unfamilar ones, rendering AI-generated artifacts more detectable (Diel & Lewis, 2022).

The quality of the deepfake generation process can influence performance, for example, through resolution, noise, or anomalies. Holmes et al. (2016) found higher resolution improved detection performance for artificial faces. Low quality deepfakes are also more likely to be detected due to generation noise, hallucinations, and other artifacts. This meta-analysis does not show above-chance detection performance, indicating that the deepfakes in the included studies were of good quality. Some studies report higher accuracy (e.g., Groh et al., 2022, 2024), possibly due to high stimulus resolution or familiarity with the content (e.g., celebrities). Caution is advised in generalizing these results to other cases, as detection may vary based on specific factors. The moderating effects of generation quality, stimulus resolution, and content familarity across modalities could be investigated in future research.

This meta-analysis is limited by the reviewed papers' selective and heterogeneous reporting of results. Even when studies used two-alternative forced choice tasks to collect binominal fake/real categorization data for deepfake and real faces, several studies focused on specific analyses, such as AUC (Groh et al., 2022; Korshunov & Marcel, 2020), *d'* (Holmes et al., 2016; Nightingale & Farid, 2022), total correct identification (Cooke et al., 2024; Shen et al., 2021), and confusion matrices including hits and false alarms (Gao et al., 2023; Ha et al., 2024). However, not all studies report the values required to perform all analyses. Future meta-analyses could benefit from studies reporting hits, false alarms, and variance measures, enabling further analyses (e.g., for

ORs). The current meta-analysis had too few studies for certain subgroup analyses, such as investigating strategies to improve detection. Nevertheless, the meta-analyses on proportion, OR, and *d'* results indicate robust outcomes, with no variation in dimensions that could affect accuracy.

Another limitation is the heterogeneous methodologies used in the reviewed studies, which vary in stimulus quality, content, and other characteristics, sample size, survey questions, and experimental manipulations (e.g., awareness of deepfake presence). A random effect model controls for overall heterogeneity, reducing the effect of differences in study design and stimulus characteristics at the cost of wider variances. Although some dimensions were controlled (e.g., weighting by sample and stimulus size, excluding low-quality deepfakes), others were not. Future reviews could explore different aspects of deepfakes (e.g., artifacts, complexity, believability, setting, and familiarity) through a qualitative analysis of differences in the literature.

Of the 56 papers included in this systematic review and meta-analysis, 19 did not report sufficient data for all analyses. Thus, the number of studies included varied: total accuracies (56 studies), deepfake accuracies (39 studies), real stimulus accuracies (36 studies), OR (35 studies), and *d'* (35 studies).

The purpose of a meta-analysis is to increase statistical power beyond individual studies (Cohn & Becker, 2003). The number of studies met general recommendations for meta-analyses (Jackson & Turner, 2017; Valentine, Pigott, & Rothstein, 2010). Thanks to large sample sizes, a post-hoc power analysis (Valentine, Pigott, & Rothstein, 2010) on the smallest subgroup (deepfake audio, $k = 8$) revealed power of $1 - \beta > 0.99$, indicating a very high probability of correctly rejecting the null hypothesis. Thus, nonsignificant results cannot be attributed to low statistical power.

5. Conclusion

AI-generated content is becoming increasingly realistic, raising ethical concerns when it is used to deceive. The misuse of AI threatens public security through disinformation, propaganda, financial fraud, identity theft, and pornography. Human guesses on whether AI-generated content is fake or real are at chance levels. However, several techniques can improve human performance, and combined with human–AI collaboration, these methods can guard against deception. Combining strategies to improve human deepfake detection may be especially useful. Despite no evidence of above-chance deepfake detection, some studies reported above-chance performance; hence, caution should be taken when attempting to generalize the present results as deepfake detection performance may depend on various factors (e.g., deepfake quality).

CRediT authorship contribution statement

Alexander Diel: Writing – review & editing, Writing – original draft, Visualization, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Tania Lalgi:** Writing – review & editing, Visualization, Validation, Data curation. **Isabel Carolin Schröter:** Writing – review & editing, Validation, Data curation. **Karl F. MacDorman:** Writing – review & editing, Visualization, Validation, Methodology, Formal analysis. **Martin Teufel:** Supervision, Resources, Project administration. **Alexander Bäuerle:** Writing – review & editing, Supervision, Project administration, Conceptualization.

Funding information

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

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.

Appendix

Table A1
Summary of the studies included in this meta-analysis in alphabetical order.

Study	Stimulus type	Strategy used	Data available
Ahmed et al. (2021)	Video	Awareness	Partially available
Ask et al. (2023)	Video	No	Fully in manuscript
Barari, Lucas, and Munger (2021)	Video	No	Partially available
Boyd et al. (2023)	Image	AI support	Fully available
Bray et al. (2023)	Image	Advice	Fully in manuscript
Cartella, Cuculo, Cornia, and Cucchiara (2024)	Image	Advice	Fully available
Chein, Martinez, and Barone (2024)	Text	No	Fully in manuscript
Cooke et al. (2024)	Audio, image, video	No	Partially available
Diel et al. (2024)	Image	Feedback training	Fully available
Doss et al. (2023)	Video	No	Fully in manuscript
Fosco et al. (2022)	Video	Caricature	Partially available
Frank et al. (2023)	Audio, image, text	No	Fully available
Groh et al. (2022)	Video	AI support	Fully available
Groh et al. (2024)	Audio, text, video	No	Fully available
Gao et al. (2023)	Text	No	Fully in manuscript
Ha et al. (2024)	Image	No	Fully in manuscript
Hakam et al. (2024)	Text	No	Fully in manuscript
Han, Mitra, and Billah (2024)	Audio	No	Fully in manuscript
Hashmi et al. (2024)	Video	No	Partially available
Holmes et al. (2016)	Image	Feedback training	Partially available
Hulzebosch, Ibrahimi, and Worring (2020)	Image	Feedback training	Fully in manuscript
Jakesch, Hancock, and Naaman (2023)	Text	Financial incentive, feedback training	Contact author
Josephs et al. (2023)	Video	Caricature	Partially available
Kim et al. (2024)	Text	No	Fully in manuscript

(*continued on next page*)

Table A1 (*continued*)

Study	Stimulus type	Strategy used	Data available
Khan et al. (2023)	Video	No	Partially available
Köbis et al. (2021)	Video	Awareness, financial incentive	Partially available
Knoedler et al. (2024)	Text	No	Partially available
Korshunov and Marcel (2020)	Video	No	Partially available
Li et al. (2023)	Text	No	Fully in manuscript
Libourel, Husseini, Mirabet-Herranz, and Dugelay (2024)	Video	No	Partially available
Lovato et al. (2024)	Video	No	Fully in manuscript
Lu et al. (2024)	Image	No	Fully in manuscript
Mader, Banks, and Farid (2017)	Image	Feedback training	Partially available
Mai et al. (2023)	Audio	No	Fully in manuscript
Mittal, Sinha, Swaminathan, Collomosse, and Manocha (2023)	Video	No	Fully in manuscript
Moshel, Robinson, Carlson, and Grootswagers (2022)	Image	No	Partially available
Müller, Pizzi, and Williams (2022)	Audio	Feedback training	Partially available
Nas and De Kleijn (2024)	Video	No	Fully in manuscript
Nightingale and Farid (2022)	Image	Feedback training	Contact author
Partadiredja, Serrano, and Ljubenkov (2020)	Image, text	No	Partially available
Prasad, Hadar, Vu, and Polian (2022)	Video	No	Fully in manuscript
Preu, Jackson, and Choudhury (2022)	Image	No	Fully in manuscript
Rössler et al. (2019)	Image	No	Contact author
Salini and HariKiran (2024)	Video	No	Fully in manuscript
Silva, Khera, and Schwamm (2024)	Text	No	Fully in manuscript
Shen et al. (2021)	Image	No	Partially available
Somoray and Miller (2023)	Video	Advice	Contact author
Stadler et al. (2024)	Text	No	Fully in manuscript
Tahir et al. (2021)	Video	Advice	Fully in manuscript
Thaw, July, Wai, Goh, and Chua (2021)	Video	No	Partially available
Tucciarelli et al. (2022a	Image	Awareness	Partially available
Uchendu et al. (2023)	Text	Human support	Contact author
Vaccari and Chadwick (2020)	Video	No	Contact author

Note. "Data available" refers to the degree of availability of the data extracted from the studies. "Partially available" indicates that some data was made available in the manuscript that was used for some but not sufficient for all analyses (e.g., total accuracy but not hit and correct rejection rates); "Fully in manuscript" indicates that all relevant data was reported in the manuscript; "fully available" indicates that raw data was publicly available and used for the relevant analyses; "author contact" indicates that raw or relevant data was shared after contacting the authors.

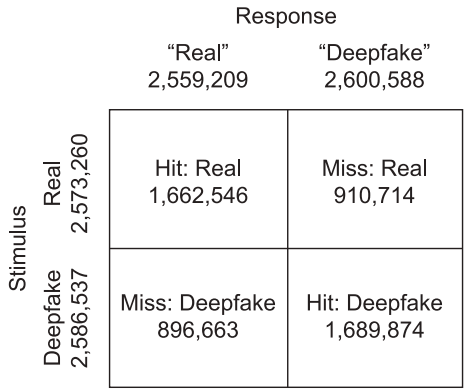

Fig. A1. Confusion matrix depicting detection performance for real and deepfake stimuli across all studies. Out of 5,159,797 responses, 3,352,420 were correct (hits) and 1,807,377 were incorrect (misses).

Data availability

The data and analysis are publicly available at https://osf.io/hnf8g/

References

Adams, Z., Osman, M., Bechlivanidis, C., & Meder, B. (2023). (Why) Is misinformation a problem? *Perspectives on Psychological Science, 18*(6), 1436–1463. https://doi.org/10.1177/17456916221141344

Ahmed, M. F. B., Miah, M. S. U., Bhowmik, A., & Sulaiman, J. B. (2021). Awareness to deepfake: A resistance mechanism to deepfake. In *Proceedings of the 2021 international Congress of advanced Technology and engineering (ICOTEN)* (pp. 1–5). IEEE.

Ahmed, M. F. B., Miah, M. S. U., Bhowmik, A., & Sulaiman, J. B. (2021). Awareness to deepfake: A resistance mechanism to deepfake. In *Proceedings of the 2021 international Congress of advanced Technology and engineering (ICOTEN)* (pp. 1–5). New York: IEEE.

Aïmeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: A review. *Social Network Analysis and Mining, 13*(1), 30. https://doi.org/10.1007/s13278-023-01028-5

Amerini, I., Ballan, L., Caldelli, R., Bimbo, A., Tongo, L. D., & Serra, G. (2013). Copy-move forgery detection and localization by means of robust clustering with J-Linkage. *Signal Processing: Image Communication, 28*, 659–669. https://doi.org/10.1016/j.image.2013.03.006

Ask, T. F., Lugo, R., Fritsch, J., Veng, K., Eck, J., Özmen, M. T., … Sütterlin, S. (2023). Cognitive flexibility but not cognitive styles influence deepfake detection skills and metacognitive accuracy. *PsyArXiv Preprints–OSF*. https://doi.org/10.31234/osf.io/a9dwe

Barari, S., Lucas, C., & Munger, K. (2021). Political deepfakes are as credible as other fake media and (sometimes) real media. *OSF Preprints, 13*. https://doi.org/10.31219/osf.io/cdfh3

Bateman, J. (2020). *Deepfakes and synthetic media in the financial system: Assessing threat scenarios*. Washington, DC: Carnegie Endowment for International Peace.

Borenstein, M., Hedges, L. V., Higgins, P. T., & Rothstein, H. (2009). *An introduction to meta-analysis*. Hoboken, NJ: Wiley.

Boyd, A., Tinsley, P., Bowyer, K., & Czaijka, A. (2023). The value of AI guidance in human examination of synthetically-generated faces. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence, 665* pp. 5930–5938). https://doi.org/10.1609/aaai.v37i5.25734

Bray, S. D., Johnson, S. D., & Kleinberg, B. (2023). Testing human ability to detect 'deepfake' images of human faces. *Journal of Cybersecurity, 9*(1), Article tyad011. https://doi.org/10.1093/cybsec/tyad011

Bundesamt für Sicherheit in der Informationstechnik. (2024). Deepfakes – Gefahren und Gegenmaßnahmen. Retrieved July 2024 from https://www.bsi.bund.de/DE/Themen/Unternehmen-und-Organisationen/Informationen-und-Empfehlungen/Kuenstliche-Intelligenz/Deepfakes/deepfakes_node.html.

Bundesministerium Inneres. (2024). *Deepfakes*. Vienna, Austria: Bundesministerium für Inneres. Retrieved July 2024 from: https://www.bmi.gv.at/magazin/2024_03_04/04_Deepfakes.aspx.

Caldwell, M., Andrews, J. T. A., Tanay, T., & Griffin, L. D. (2020). AI-enabled future crime. *Crime Science, 9*, 14. https://doi.org/10.1186/s40163-020-00123-8

Campbell, C., Plangger, K., Sands, S., & Kietzmann, J. (2021). Preparing for an era of deepfakes and AI-generated ads: A framework for understanding responses to manipulated advertising. *Journal of Advertising, 51*(1), 22–38. https://doi.org/10.1080/00913367.2021.1909515

Canadian Security Intelligence Service. (2023). Deepfakes: A real threat to a Canadian future. Retrieved from (July 2024) https://www.canada.ca/en/security-intelligence-service/corporate/publications/the-evolution-of-disinformation-a-deepfake-future/deepfakes-a-real-threat-to-a-canadian-future.html.

Cartella, G., Cuculo, V., Cornia, M., & Cucchiara, R. (2024). Unveiling the truth: Exploring human gaze patterns in fake images. *IEEE Signal Processing Letters, 31*, 820–824. https://doi.org/10.1109/LSP.2024.3375288

Chadha, A., Kumar, V., Kashyap, S., & Gupta, M. (2021). Deepfake: An overview. In P. K. Singh, S. T. Wierzchoń, S. Tanwar, M. Ganzha, & J. J. P. C. Rodrigues (Eds.), *Lecture notes in networks and systems: Vol. 203. Proceedings of second international conference on computing, communications, and cyber-security*. Singapore: Springer. https://doi.org/10.1007/978-981-16-0733-2_39.

Chein, J., Martinez, S., & Barone, A. (2024). Can human intelligence safeguard against artificial intelligence? Exploring individual differences in the discernment of human from AI texts. *Research Square Preprint*. https://doi.org/10.21203/rs.3.rs-4277893/v1. April 29.

Cohn, L. D., & Becker, B. J. (2003). How meta-analysis increases statistical power. *Psychological Methods, 8*(3), 243. https://doi.org/10.1037/1082-989X.8.3.243

Cooke, D., Edwards, A., Barkoff, S., & Kelly, K. (2024). As good as a coin toss: Human detection of AI-generated images, videos, audio, and audiovisual stimuli. *arXiv Preprint*. https://doi.org/10.48550/arXiv.2403.16760

Dai, Z., & MacDorman, K. F. (2021). Creepy, but persuasive: In a virtual consultation, physician bedside manner, rather than the uncanny valley, predicts adherence. *Frontiers in Virtual Reality, 2*, 1–18. https://doi.org/10.3389/frvir.2021.739038, 739038.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials, 7*(3), 177–188.

Diel, A., & Lewis, M. (2022). Familiarity, orientation, and realism increase face uncanniness by sensitizing to facial distortions. *Journal of Vision, 22*(4), 14. https://doi.org/10.1167/jov.22.4.14

Diel, A., & Lewis, M. (2024). Deviation from typical organic voices best explains a vocal uncanny valley. *Computers in Human Behavior Reports, 14*, Article 100430. https://doi.org/10.1016/j.chbr.2024.100430

Diel, A., Teufel, M., & Bäuerle, A. (2024). Inability to detect deepfakes: Deepfake detection training improves detection accuracy, but increases emotional distress and reduces self-efficacy. Preprint. Retrieved from: https://osf.io/preprints/osf/muwnj.

Doss, C., Mondschein, J., Shu, D., Wolfson, T., Kopecky, D., Fitton-Kane, V. A., et al. (2023). Deepfakes and scientific knowledge dissemination. *Scientific Reports, 13*(1), Article 13429. https://doi.org/10.1038/s41598-023-39944-3

Eberl, A., Kühn, J., & Wolbring, T. (2022). Using deepfakes for experiments in the social sciences: A pilot study. *Frontiers in Sociology, 7*, Article 907199. https://doi.org/10.3389/fsoc.2022.907199

Elali, F. R., & Rachid, L. N. (2023). AI-generated research paper fabrication and plagiarism in the scientific community. *Patterns, 4*(3), Article 100706. https://doi.org/10.1016/j.patter.2023.100706

Fagni, T., Falchi, F., Gambini, M., Martella, A., & Tesconi, M. (2021). TweepFake: About detecting deepfake tweets. *PLoS One, 16*(5), Article e0251415. https://doi.org/10.1371/journal.pone.0251415

Farid, H. (2022). Creating, using, misusing, and detecting deep fakes. *Journal of Online Trust and Safety, 1*(4). https://doi.org/10.54501/jots.v1i4.56

Fink, G. (2019). Adversarial artificial intelligence: State of the malpractice. *Journal of Information Warfare, 18*(4), 1–23. https://www.jstor.org/stable/26894691.

Fosco, C., Josephs, E., Andonian, A., Lee, A., Wang, X., & Oliva, A. (2022). Deepfake Caricatures: Amplifying attention to artifacts increases deepfake detection by humans and machines. *arXiv Preprint*. https://doi.org/10.48550/arXiv.2206.00535

Frank, J., Herbert, F., Ricker, J., Schönherr, L., Eisenhofer, T., Fischer, A., … Holz, T. (2023). A representative study on human detection of artificially generated media across countries. *arXiv Preprint*. https://doi.org/10.48550/arXiv.2312.05976

Freeman, M. F., & Tukey, J. W. (1950). Transformations related to the angular and the square root. *The Annals of Mathematical Statistics, 21*(4), 607–611. http://www.jstor.org/stable/2236611.

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., et al. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *NPJ Digital Medicine, 6*(1), 75. https://doi.org/10.1038/s41746-023-00819-6

Groh, M., Epstein, Z., Firestone, C., & Picard, R. W. (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

Groh, M., Sankaranarayanan, A., Singh, N., Kim, D. Y., Lippman, A., & Picard, R. (2024). Human detection of political speech deepfakes across transcripts, audio, and video. *Nature Communications, 15*(1), 7629. https://doi.org/10.1038/s41467-024-51998-z

Ha, A. Y. J., Passananti, J., Bhaskar, R., Shan, S., Southen, R., Zheng, H., et al. (2024). Organic or diffused: Can we distinguish human art from AI-generated images? *arXiv Preprint*. https://doi.org/10.48550/arXiv.2402.03214

Hadland, A., Cambell, D., & Lambert, P. (2015). *The state of news photography: The lives and livelihoods of photojournalists in the digital age*. Oxford: UK: Reuters Institute for the Study of Journalism. https://doi.org/10.60625/risj-hjmq-sh17

Hakam, H. T., Prill, R., Korte, L., Lovreković, B., Ostojić, M., Ramadanov, N., et al. (2024). Human-written vs AI-generated texts in orthopedic academic literature: Comparative qualitative analysis. *JMIR Formative Research, 8*, Article e52164. https://doi.org/10.2196/52164

Hamed, S. K., Ab Aziz, M. J., & Yaakub, M. R. (2023). A review of fake news detection approaches: A critical analysis of relevant studies and highlighting key challenges associated with the dataset, feature representation, and data fusion. *Heliyon, 9*(10), Article e20382. https://doi.org/10.1016/j.heliyon.2023.e20382

Han, C., Mitra, P., & Billah, S. M. (2024). Uncovering human traits in determining real and spoofed audio: Insights from blind and sighted individuals. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (pp. 1–14). https://doi.org/10.1145/3613904.3642817

Hao, K. (2021). Deepfake porn is ruining women's lives. Now the law may finally ban it. *MIT Technology Review*. Retrieved July 2024 from: https://www.technologyreview.com/2021/02/12/1018222/deepfake-revenge-porn-coming-ban/.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering, 21*(9), 1263–1284.

Heidari, A., Navimipour, N. J., Dag, H., & Unal, M. (2024). Deepfake detection using deep learning methods: A systematic and comprehensive review. *WIREs Data Mining and Knowledge Discovery, 14*(2), Article e1520. https://doi.org/10.1002/widm.1520

Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses. *BMJ, 327*(7414), 557–560.

Holmes, O., Banks, M. S., & Farid, H. (2016). Assessing and improving the identification of computer-generated portraits. *ACM Transactions on Applied Perception, 13*(2), 1–12. https://doi.org/10.1145/2871714

Homeland Security. (2022). Increasing threat of deepfake identities. Retrieved July 2024 from https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf.

Hulzebosch, N., Ibrahimi, S., & Worring, M. (2020). Detecting CNN-generated facial images in real-world scenarios. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. https://doi.org/10.1109/CVPRW50498.2020.00329

Ibrahim, H., Liu, F., Asim, R., Battu, B., Benabderrahmane, S., Alhafni, B., et al. (2023). Author correction: Perception, performance, and detectability of conversational artificial intelligence across 32 university courses. *Scientific Reports, 13*(1), Article 17101. https://doi.org/10.1038/s41598-023-43998-8

Jackson, D., & Turner, R. (2017). Power analysis for random-effects meta-analysis. *Research Synthesis Methods, 8*(3), 290–302. https://doi.org/10.1002/jrsm.1240

Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences, 120*(11), Article e2208839120. https://doi.org/10.1073/pnas.2208839120

Josephs, E., Fosco, C., & Oliva, A. (2023). Artifact magnification on deepfake videos increases human detection and subjective confidence. *Journal of Vision, 23*(9), 5327. https://doi.org/10.1167/jov.23.9.5327

Judge, S., & Hayton, N. (2022). Voice banking for individuals living with mnd: A service review. *Technology and Disability, 34*(2), 113–122. https://doi.org/10.3233/tad-210366

Juefei-Xu, F., Wang, R., Huang, Y., Guo, Q., Ma, L., & Liu, Y. (2022). Countering malicious deepfakes: Survey, battleground, and horizon. *International Journal of Computer Vision, 130*(7), 1678–1734. https://doi.org/10.1007/s11263-022-01606-8

Katanich, D. (2024). It's a scam! How deepfakes and voice cloning tap into your cash. *EuroNews*. Retrieved July 2024 from https://www.euronews.com/business/2024/04/10/its-a-scam-how-deepfakes-and-voice-cloning-taps-into-your-cash.

Keya, A. J., Shajeeb, H. H., Rahman, M. S., & Mridha, M. F. (2023). FakeStack: Hierarchical Tri-BERT-CNN-LSTM stacked model for effective fake news detection. *PLoS One, 18*(12), Article e0294701. https://doi.org/10.1371/journal.pone.0294701

Khan, M. R., Naeem, S., Tariq, U., Dhall, A., Khan, M. N. A., Al Shargie, F., et al. (2023). In *Exploring neurophysiological responses to cross-cultural deepfake videos. Companion Publication of the* (pp. 41–45). https://doi.org/10.1145/3610661.361714

Khanjani, Z., Watson, G., & Janeja, V. P. (2023). Audio deepfakes: A survey. *Frontiers in Big Data, 5*, Article 1001063. https://doi.org/10.3389/fdata.2022.1001063

Kim, H. J., Yang, J. H., Chang, D. G., Lenke, L. G., Pizones, J., Castelein, R., … Suk, S. I. (2024). Assessing the reproducibility of the structured abstracts generated by ChatGPT and Bard compared to human-written abstracts in the field of spine surgery: Comparative analysis. *Journal of Medical Internet Research, 26*, Article e52001. https://doi.org/10.2196/52001

Knoedler, S., Sofo, G., Kern, B., Frank, K., Cotofana, S., von Isenburg, S., … Alfertshofer, M. (2024). Modern machiavelli? The illusion of ChatGPT-generated patient reviews in plastic and aesthetic surgery based on 9000 review classifications. *Journal of Plastic, Reconstructive & Aesthetic Surgery, 88*, 99–108. https://doi.org/10.1016/j.bjps.2023.10.119

Köbis, N. C., Doležalová, 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 Preprints*. https://doi.org/10.48550/arXiv.2009.03155

Li, Y., Li, Q., Cui, L., Bi, W., Wang, L., Yang, L., … Zhang, Y. (2023). Deepfake text detection in the wild. *arXiv Preprints*. arXiv:2305.13242.

Libourel, A., Husseini, S., Mirabet-Herranz, N., & Dugelay, J. L. (2024). A case study on how beautification filters can fool deepfake detectors. In *IWBF 2024, 12th IEEE International Workshop on Biometrics and Forensics*, Article 10593932. https://doi.org/10.1109/IWBF62628.2024

Lovato, J., St-Onge, J., Harp, R., Salazar Lopez, G., Rogers, S. P., Haq, I. U., … Onaolapo, J. (2024). Diverse misinformation: Impacts of human biases on detection of deepfakes on networks. *NPJ Complexity, 1*(1), 5. https://doi.org/10.1038/s44260-024-00006-y

Lu, Z., Huang, D., Bai, L., Qu, J., Wu, C., Liu, X., et al. (2024). Seeing is not always believing: Benchmarking human and model perception of AI-generated images. *Advances in Neural Information Processing Systems, 1105*, 25435–25447. https://doi.org/10.5555/3666122.3667227

Lyu, S. (2020). In *Deepfake detection: Current challenges and next steps* (pp. 1–6). IEEE. https://doi.org/10.1109/ICMEW46912.2020.9105991.

Macmillan, N. A. (2002). Signal detection theory. In H. Pashler (Ed.), *Stevens' handbook of experimental psychology: Methodology in experimental psychology* (4th ed., Vol. 4, pp. 43–90). Hoboken, New Jersey: Wiley.

Mader, B., Banks, M. S., & Farid, H. (2017). Identifying computer-generated portraits: The importance of training and incentives. *Perception, 46*(9), 1062–1076. https://doi.org/10.1177/0301006617713633

Mai, K. T., Bray, S., Davies, T., & Griffin, L. D. (2023). Warning: Humans cannot reliably detect speech deepfakes. *PLoS One, 18*(8), Article e0285333. https://doi.org/10.1371/journal.pone.0285333

Májovský, M., Černý, M., Kasal, M., Komarc, M., & Netuka, D. (2023). Artificial intelligence can generate fraudulent but authentic-looking scientific medical articles: Pandora's box has been opened. *Journal of Medical Internet Research, 25*, Article e46924. https://doi.org/10.2196/46924

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22*(4), 719–748. https://doi.org/10.1093/jnci/22.4.719

Moshel, M. L., Robinson, A. K., Carlson, T. A., & Grootswagers, T. (2022). Are you for real? Decoding realistic AI-generated faces from neural activity. *Vision Research, 199*, Article 108079. https://doi.org/10.1016/j.visres.2022.108079

Müller, N. M., Pizzi, K., & Williams, J. (2022). Human perception of audio deepfakes. In *Proceedings of the 1st International Workshop on Deepfake Detection for Audio Multimedia* (pp. 85–91).

Mittal, T., Sinha, R., Swaminathan, V., Collomosse, J. P., & Manocha, D. (2023). Video manipulations beyond faces: A dataset with human-machine analysis. *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*. https://doi.org/10.1109/WACVW58289.2023.00071

Nas, E., & De Kleijn, R. (2024). Conspiracy thinking and social media use are associated with ability to detect deepfakes. *Telematics and Informatics, 87*, Article 102093. https://doi.org/10.1016/j.tele.2023.102093

Nightingale, S. J., & Farid, H. (2022). AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences, 119*(8), Article e2120481119. https://doi.org/10.1073/pnas.2120481119

Nightingale, S. J., Wade, K. A., & Watson, D. G. (2017). Can people identify original and manipulated photos of real-world scenes? *Cognitive Research, 2*(30), 1–21. https://doi.org/10.1186/s41235-017-0067-2

Odri, G. A., & Yoon, D. J. Y. (2023). Detecting generative artificial intelligence in scientific articles: Evasion techniques and implications for scientific integrity. *Orthopaedics and Traumatology: Surgery & Research, 109*(8), Article 103706. https://doi.org/10.1016/j.otsr.2023.103706

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., … Moher, D. (2021). Updating guidance for reporting systematic reviews: Development of the PRISMA 2020 statement. *Journal of Clinical Epidemiology, 134*, 103–112. https://doi.org/10.1016/j.jclinepi.2021.02.003

Panda Security. (2024). Deepfake pornography explosion. Retrieved July 2024 from https://www.pandasecurity.com/en/mediacenter/deepfake-pornography-explosion/.

Partadiredja, R. A., Serrano, C. E., & Ljubenkov, D. (2020). AI or human: The socio-ethical implications of AI-generated media content. In *Proceedings of the 13th CMI Conference on Cybersecurity and privacy (CMI). Digital transformation: Potentials and challenges (51275)* (pp. 1–6). New York, NY: IEEE. https://doi.org/10.1109/CMI51275.2020.9322673.

Piva, A. (2013). An overview on image forensics. *ISRN Signal Processing, 496701*, 1–22. https://doi.org/10.1155/2013/496701

Popkov, A. A., & Barrett, T. S. (2024). AI vs. academia: Experimental study on AI text detectors' accuracy in behavioral health academic writing. *Accountability in Research*, 1–17. https://doi.org/10.1080/08989621.2024.2331757

Prasad, S. S., Hadar, O., Vu, T., & Polian, I. (2022). Human vs. automatic detection of deepfake videos over noisy channels. In *IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1–6). New York: IEEE. https://doi.org/10.1109/ICME52920.2022.9859954.

Preu, E., Jackson, M., & Choudhury, N. (2022). Perception vs. reality: Understanding and evaluating the impact of synthetic image deepfakes over college students. In *IEEE 13th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)* (pp. 547–553). New York: IEEE. https://doi.org/10.1109/UEMCON54665.2022.9965697.

Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). Deepfake detection: A systematic literature review. *IEEE Access, 10*, 25494–25513. https://doi.org/10.1109/ACCESS.2022.3154404

Rashidi, H. H., Fennell, B. D., Albahra, S., Hu, B., & Gorbett, T. (2023). The ChatGPT conundrum: Human-generated scientific manuscripts misidentified as AI creations by AI text detection tool. *Journal of Pathology Informatics, 14*, Article 100342. https://doi.org/10.1016/j.jpi.2023.100342

Ray, S. (2020). Bot generated fake nudes of over 100,000 women without their knowledge, says report. *Forbes*. Retrieved July 2024 from: https://www.forbes.com/sites/siladityaray/2020/10/20/bot-generated-fake-nudes-of-over-100000-women-without-their-knowledge-says-report/.

Robertson, D. J., Mungall, A., Watson, D. G., Wade, K. A., Nightingale, S. J., & Butler, S. (2018). Detecting morphed passport photos: A training and individual differences approach. *Cognitive Research, 3*, 27. https://doi.org/10.1186/s41235-018-0113-8

Rocha, A., Scheirer, W., Boult, T., & Goldenstein, S. (2011). Vision of the unseen: Current trends and challenges in digital image and video forensics. *ACM Computing Surveys, 43*(4), 1–42. https://doi.org/10.1145/1978802.1978805

Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Niessner, M. (2019). FaceForensics++: Learning to detect manipulated facial images. *IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, Korea. https://doi.org/10.1109/ICCV.2019.00009

Rupapara, V., Rustam, F., Amaar, A., Washington, P. B., Lee, E., & Ashraf, I. (2021). Deepfake tweets classification using stacked Bi-LSTM and words embedding. *PeerJ Computer Science, 7*, e745. https://doi.org/10.7717/peerj-cs.745

Salini, Y., & HariKiran, J. (2024). Deepfake videos detection using crowd computing. *International Journal of Information Technology, 16*, 4547–4564. https://doi.org/10.1007/s41870-023-01494-2

Sanders, J. G., Ueda, Y., Yoshikawa, S., & Jenkins, R. (2019). More human than human: A Turing test for photographed faces. *Cognitive Research, 4*(1), 43. https://doi.org/10.1186/s41235-019-0197-9

Schetinger, V., Oliveira, M. M., da Silva, R., & Carvalho, T. J. (2017). Humans are easily fooled by digital images. *Computers & Graphics, 68*, 142–151. https://doi.org/10.1016/j.cag.2017.08.010

Seow, J. W., Lim, M. K., Phan, R. C. W., & Liu, J. K. (2022). A comprehensive overview of deepfake: Generation, detection, datasets, and opportunities. *Neurocomputing, 513*, 351–371. https://doi.org/10.1016/j.neucom.2022.09.135

Shen, B., Richard Webster, B., O'Toole, A., Bowyer, K., & Scheirer, W. J. (2021). A study on the human perception of synthetic faces. *16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), Jodhpur, India, 1–8*. https://doi.org/10.1109/FG52635.2021.9667066

Silva, G. S., Khera, R., & Schwamm, L. H. (2024). Reviewer experience detecting and judging human versus artificial intelligence content: The stroke journal essay contest. *Stroke, 55*, 10. https://doi.org/10.1161/STROKEAHA.124.04501

Simonite, T. (2021). It began as an AI-fueled dungeon game. It got much darker. *Wired*. Retrieved July 2024 from https://www.wired.com/story/ai-fueled-dungeon-game-got-much-darker/.

Somoray, K., & Miller, D. J. (2023). Providing detection strategies to improve human detection of deepfakes: An experimental study. *Computers in Human Behavior, 149*, Article 107917. https://doi.org/10.1016/j.chb.2023.107917

Stadler, R. D., Sudah, S. Y., Moverman, M. A., Denard, P. J., Duralde, X. A., Garrigues, G. E., … Menendez, M. E. (2024). Identification of ChatGPT-generated abstracts within shoulder and elbow surgery poses a challenge for reviewers. *Arthroscopy*. https://doi.org/10.1016/j.arthro.2024.06.045

Stroebel, L., Llewellyn, M., Hartley, T., Ip, T. S., & Ahmed, M. (2023). A systematic literature review on the effectiveness of deepfake detection techniques. *Journal of Cyber Security Technology, 7*(2), 83–113. https://doi.org/10.1080/23742917.2023.2192888

Strupp, C. (2019). Fraudsters used AI to mimic CEO's voice in unusual cybercrime case. *WSJ Pro Cybersecurity*. retrieved in May 2024 https://www.wsj.com/articles/fraudsters-use-ai-to-mimic-ceos-voice-in-unusual-cybercrime-case-11567157402.

Sutton, A. J. (2009). Publication bias. In H. Cooper, L. V. Hedges, & J. C. Valentine (Eds.), *The handbook of research synthesis and meta-analysis* (2nd ed., pp. 435–452). New York, NY: Russell Sage Foundation.

Suwajanakorn, S., Seitz, S. M., & Kemelmacher-Shlizerman, I. (2017). Synthesizing Obama. *ACM Transactions on Graphics, 36*, 1–13. https://doi.org/10.1145/3072959.3073640

Tahir, R., Batool, B., Jamshed, H., Jameel, M., Anwar, M., Ahmed, F., … Zaffar, M. F. (2021). Seeing is believing: Exploring perceptual differences in deepfake videos. In *Proceedings of the 2021 CHI conference on human factors in computing systems* (Vol. 174, pp. 1–16). https://doi.org/10.1145/3411764.3445699

Tharwat, A. (2021). Classification assessment methods. *Applied Computing and Informatics, 17*(1), 168–192. https://doi.org/10.1016/j.aci.2018.08.003

Thaw, N. N., July, T., Wai, A. N., Goh, D. H. L., & Chua, A. Y. (2021). How are deepfake videos detected? An initial user study. HCI International 2021 posters: 23rd HCI International Conference, July 24–29, 2021, proceedings. In , *23. Part I* (pp. 631–636). Springer. https://doi.org/10.1007/978-3-030-78635-9_80.

Tucciarelli, R., Vehar, N., Chandaria, S., & Tsakiris, M. (2022). On the realness of people who do not exist: The social processing of artificial faces. *iScience, 25*(12), Article 105441. https://doi.org/10.1016/j.isci.2022.105441

Twomey, J., Ching, D., Aylett, M. P., Quayle, M., Linehan, C., & Murphy, G. (2023). Do deepfake videos undermine our epistemic trust? A thematic analysis of tweets that discuss deepfakes in the Russian invasion of Ukraine. *PLoS One, 18*(10), Article e0291668. https://doi.org/10.1371/journal.pone.0291668

Uchendu, A., Lee, J., Shen, H., Le, T., Huang, T.-H. K., & Lee, D. (2023). Does human collaboration enhance the accuracy of identifying LLM-generated deepfake texts? *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 11*(1), 163–174. https://doi.org/10.1609/hcomp.v11i1.27557

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media + Society, 6*(1), Article 2056305120903408. https://doi.org/10.1177/2056305120903408

Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How many studies do you need? A primer on statistical power for meta-analysis. *Journal of Educational and Behavioral Statistics, 35*(2), 215–247. https://doi.org/10.3102/1076998609346961

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software, 36*(3), 1–48. https://doi.org/10.18637/jss.v036.i03

Weiss, M. (2019). Deepfake bot submissions to federal public comment websites cannot be distinguished from human submissions. *Technology Science, 2019121801*. https://techscience.org/a/2019121801/.

Whittaker, L., Mulcahy, R., Letheren, K., Kietzmann, J., & Russell-Bennett, R. (2023). Mapping the deepfake landscape for innovation: A multidisciplinary systematic review and future research agenda. *Technovation, 125*, Article 102784. https://doi.org/10.1016/j.technovation.2023.102784

Whyte, C. (2020). Deepfake news: AI-enabled disinformation as a multi-level public policy challenge. *Journal of Cyber Policy, 5*(2), 199–217. https://doi.org/10.1080/23738871.2020.1797135

Winnard, N. (2024). *The rise of deepfakes in schools*. TIE Online. Retrieved from https://www.tieonline.com/article/3632/the-rise-of-deepfakes-in-schools.

Yu, P., Xia, Z., Fei, J., & Lu, Y. (2021). A survey on deepfake video detection. *IET Biometrics, 10*(6), 607–624. https://doi.org/10.1049/bme2.12031

Zalake, M. (2023). Doctors' perceptions of using their digital twins in patient care. *Scientific Reports, 13*, Article 21693. https://doi.org/10.1038/s41598-023-48747-5

Zhou, X., Ling, Z.-H., & King, S. (2020). The blizzard challenge 2020. *Proceedings of the Joint Workshop for the Blizzard Challenge and Voice Conversion Challenge*, 1–18. https://doi.org/10.21437/VCCBC.2020-1, 2020.

Zotov, S., Dremliuga, R., Borshevnikov, A., & Krivosheeva, K. (2020). Deepfake detection algorithms: A meta-analysis. In *Proceedings of the 2020 2nd symposium on signal processing systems* (pp. 43–48). https://doi.org/10.1145/3421515.3421532