Wiley
Human Behavior and Emerging Technologies
Volume 2025, Article ID 1833228, 15 pages
https://doi.org/10.1155/hbe2/1833228

WILEY

Review Article

Human Performance in Deepfake Detection: A Systematic Review

Klaire Somoray, Dan J. Miller, and Mary Holmes

James Cook University College of Healthcare Sciences, Townsville City, Australia

Correspondence should be addressed to Klaire Somoray; klaire.somoray@jcu.edu.au

Received 23 September 2024; Revised 25 June 2025; Accepted 7 July 2025

Academic Editor: Rudra Bhandari

Deepfakes refer to a wide range of computer-generated synthetic media, in which a person's appearance or likeness is altered to resemble that of another. This systematic review is aimed at providing an overview of the existing research into people's ability to detect deepfakes. Five databases (IEEE, ProQuest, PubMed, Web of Science, and Scopus) were searched up to December 2023. Studies were included if they (1) were an original study; (2) were reported in English; (3) examined people's detection of deepfakes; (4) examined the influence of an intervention, strategy, or variable on deepfake detection; and (5) reported relevant data needed to evaluate detection accuracy. Forty independent studies from 30 unique records were included in the review. Results were narratively summarized, with key findings organized based on the review's research questions. Studies used different performance measures, making it difficult to compare results across the literature. Detection accuracy varies widely, with some studies showing humans outperforming AI models and others indicating the opposite. Detection performance is also influenced by person-level (e.g., cognitive ability, analytical thinking) and stimuli-level factors (e.g., quality of deepfake, familiarity with the subject). Interventions to improve people's deepfake detection yielded mixed results. Humans and AI-based detection models focus on different aspects when detecting, suggesting a potential for human–AI collaboration. The findings highlight the complex interplay of factors influencing human deepfake detection and the need for further research to develop effective strategies for deepfake detection.

Keywords: cognitive processes; decision-making; deepfakes; human–AI collaboration; human deepfake detection

1. Introduction

Deepfakes encompass a wide range of computer-generated synthetic media, in which a person's appearance or likeness is altered to resemble that of another. The term itself combines the words *deep* (from deep learning) and *fake*, denoting that artificial intelligence (AI) techniques are used to produce hyperrealistic fake media.

The term itself has been around since 2017, when a Reddit user named u/deepfakes shared their created videos on the platform. This led to the formation of a hobbyist community built around the (now banned) subreddit r/deepfakes [1]. Although digital manipulation of images is not new, deepfakes alarmed the public due to the scale, scope, and sophistication of the resulting media and the ease with which they can be produced. Recent software developments (e.g., DeepFaceLab) and mobile apps (e.g., ZAO and FaceApp) enable users to create deepfakes with minimal programming or technological expertise. Concerns about potential malicious uses of this technology prompted research into the public's ability to detect deepfakes (e.g., [2–4]). This study, therefore, is aimed at providing a comprehensive review of the existing research into human detection of deepfakes.

Facial manipulation methods can be traced as far back as 1997 with the Video Rewrite Program [5]. The program was designed for dubbing movies using computer vision and morphing techniques. Computer vision was used to follow the movements of the speaker's mouth while morphing was used to create a new video in which the mouth movements synced to a different audio.

Since then, numerous automated facial and speech manipulation technologies have evolved. Within the facial manipulation literature, techniques used for manipulations can be categorized into five types: face swap, face morphing, lip-syncing, retouching, and face synthesis [1, 6, 7]. These

different approaches are visually summarized in Figure 1. Face swapping involves the replacement of one person's face with that of another. Face swapping can also include swapping expressions (e.g., angry to sad). Face morphing combines two or more faces to create a result that reflects each contributor's characteristics to varying degrees [7]. Lip-syncing (also known as audio or text-to-video conversion) is the process of creating video clips by integrating content from previous videos with new audio or text input [8]. Mouth movements are then aligned with new audio. Retouching includes alterations of facial features, such as makeup application or changing one's eye color, usually to enhance visual appearance [7]. Finally, face synthesis generates entirely new faces using a multitude of face images, resulting in the creation of faces that do not exist in reality [1].

AI-synthesized audio manipulation is another form of deepfakery. This technology replicates a target's voice to produce fabricated speech via text-to-speech synthesis or voice conversion [8]. Audio manipulations can be particularly challenging to detect due to the lack of visual cues to draw from [2].

Different techniques produce output of varying quality. For instance, within the DeepFake Detection Challenge (DFDC) [9] (Figure 2a) and Celeb-DF v2 [10] datasets (Figure 2b), a range of quality can be observed.

Face swapping, morphing, and lip-syncing have more potential for malicious use (as they involve replacing the image of one person with the likeness of an existing person), in contrast to retouching and face synthesis. Accordingly, the current review focuses on human detection of video deepfakes generated using face swapping, morphing, or lip-syncing techniques, as well as audio deepfakes in which an existing person's voice is imitated.

Existing studies and reviews on deepfake detection have predominantly concentrated on AI-based detection methods [1, 11]. These methods display a wide range of accuracy, from approximately 60% to 100%. Accuracy ratings are influenced by factors including, but not limited to, the specific dataset from which stimulus videos are taken (stimulus dataset) and the particular algorithm applied in the detection process [11]. Despite the existence of these technologies, deploying these techniques on a large scale requires significant computational resources. The implementation of AI-based detection on social media platforms is yet to be seen. Consequently, media consumers must rely on their own judgments when discerning whether the content they are consuming is real or fabricated.

In the same way that AI tools for spotting deepfakes show varying levels of accuracy, we expect people's accuracy levels to also vary depending on a range of factors. It can be argued that humans are more susceptible to errors in spotting deepfakes due to inherent cognitive biases—biases not present in AI-based detection technologies. However, some researchers argue for the "wisdom of the crowd" concerning human performance in deepfake detection [2]. For instance, studies have shown that under some circumstances people can surpass certain computer models in the detection of deepfakes [2]. This superior human performance has been attributed to our specialized ability to holistically visually process faces [2]. Deepfakes, being computer-generated synthetic media, often contain unintended visual "artefacts." These artefacts, which may not be salient enough for computer vision, might be perceptible to humans due to our ability to process faces holistically.

Furthermore, it is important to understand the factors that influence people's ability to detect deepfakes. There are concerns that deepfakes pose a significant threat to the trustworthiness of the information we consume—or even "reality" so to speak [12]. By understanding how people detect deepfakes, media organizations can develop better strategies to authenticate content and restore public trust. Research into human detection of deepfakes can also complement AI-based detection technologies. Understanding human strengths and weaknesses can lead to the development of hybrid detection systems that leverage both AI efficiency and human intuition.

1.1. The Current Study. While technical reviews have examined algorithmic approaches to deepfake detection (e.g., [1, 11]), our systematic review comprehensively synthesizes research specifically focusing on human detection capabilities. This paper centralizes the human element, systematically analyzing factors that enhance or impede human detection accuracy across different contexts. This human-centered approach addresses a critical gap in the literature, particularly important given that end users—not algorithms—are often the first line of defense against deepfake misinformation in everyday media consumption. Thus, the current paper is aimed at providing a comprehensive review of the literature on the human detection of deepfakes.

Specifically, the systematic review is aimed at addressing the following research questions:

- RQ1: What existing studies have examined people's ability to detect deepfakes?
- RQ2: What interventions or variables influence people's ability to detect deepfakes?

2. Methods

2.1. Search Strategy. This review was carried out according to the preferred reporting items for systematic reviews and meta-analyses (PRISMA) guidelines ([13]; see Table S1 for the checklists). The review was registered on OSF (https://osf.io/admgn/). The registration followed the PROSPERO template and included research questions, databases, inclusion/exclusion criteria, data extraction, and synthesis strategies.

A comprehensive search of five electronic databases (IEEE, ProQuest, PubMed, Web of Science, and Scopus) was conducted in December 2022 (see Table S2 for the full search strategy). We ran an updated search in December 2023 to identify any new literature. To reduce the possibility of publication bias, unpublished works and dissertations were included. While no language restriction was applied to the search, articles written in languages other than English were excluded. No time limits were imposed. Our search strategy included keywords relating to *deepfakes* and people (*subject* or *participant*). The articles

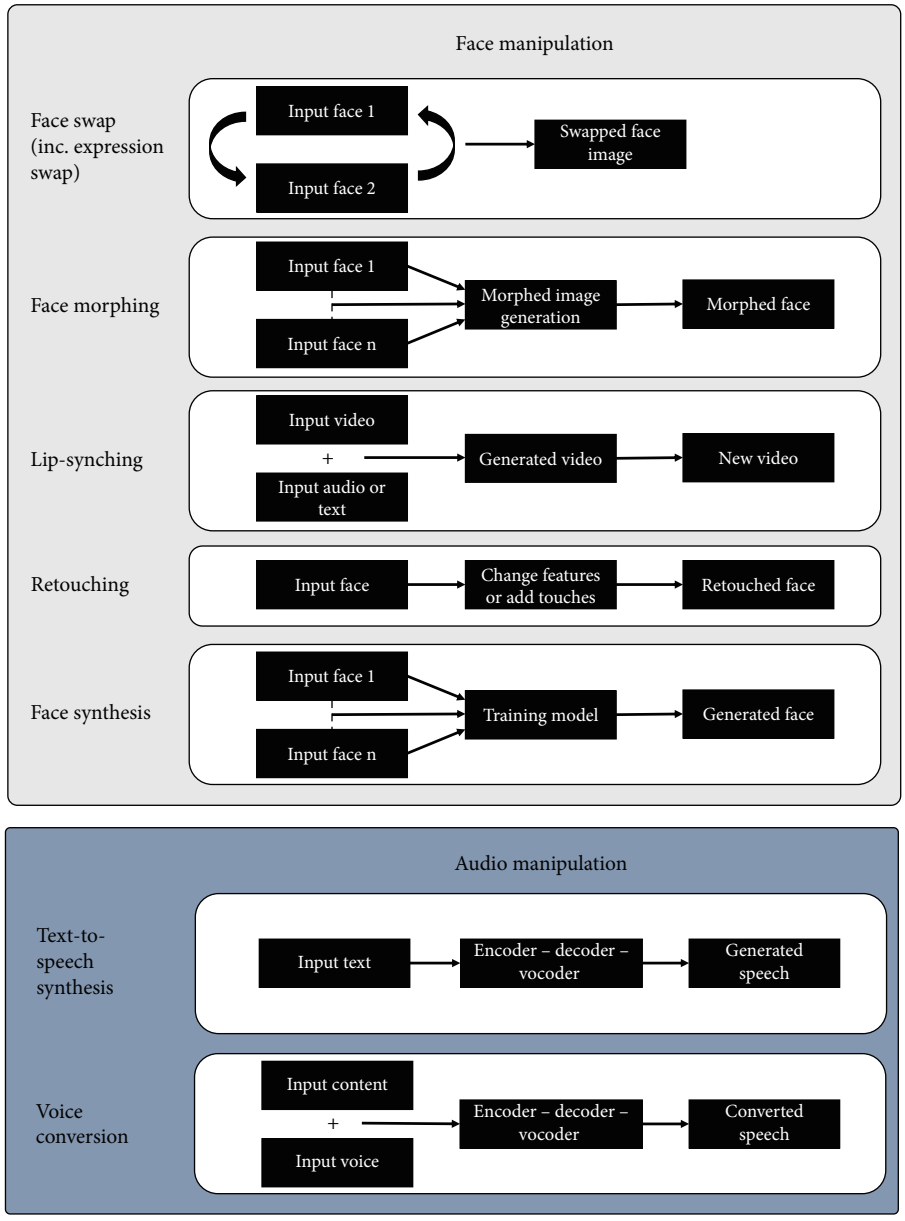

Figure 1: Different techniques used for face and speech manipulation.

included in the data extraction were searched for forward and backward citations using SpiderCite [14].

2.2. Selection Process. Rayyan.ai was used for the screening process, which involved two phases: (1) title and abstract review and (2) full-text review. For both phases, two reviewers independently screened the identified studies, and any discrepancies between reviewers were resolved through a third reviewer.

As per the registration, we included studies that met the following criteria: the record should (1) be an original study; (2) be reported in English; (3) have examined people's detection of deepfakes; (4) have examined the influence of an intervention, strategy, or variable on deepfake detection, and (5) have reported relevant data needed to evaluate detection accuracy. Studies were excluded if they (1) focused exclusively on AI-based detection methods without human participants, (2) examined general media literacy without specific deepfake detection tasks, (3) lacked criteria to assess detection accuracy, and (4) used synthetic stimuli not mimicking the likeness of an existing person or used stimuli subjected to retouching methods merely to enhance a person's features rather than substantively alter their identity. The final exclusion criterion was included to focus the review on the types of deepfakes which have the greatest potential for harm (i.e., those in which replicate an existing person's likeness).

While limiting our review to English-language publications potentially excludes relevant international research, our preliminary search indicated that most significant work in this field is published in English.

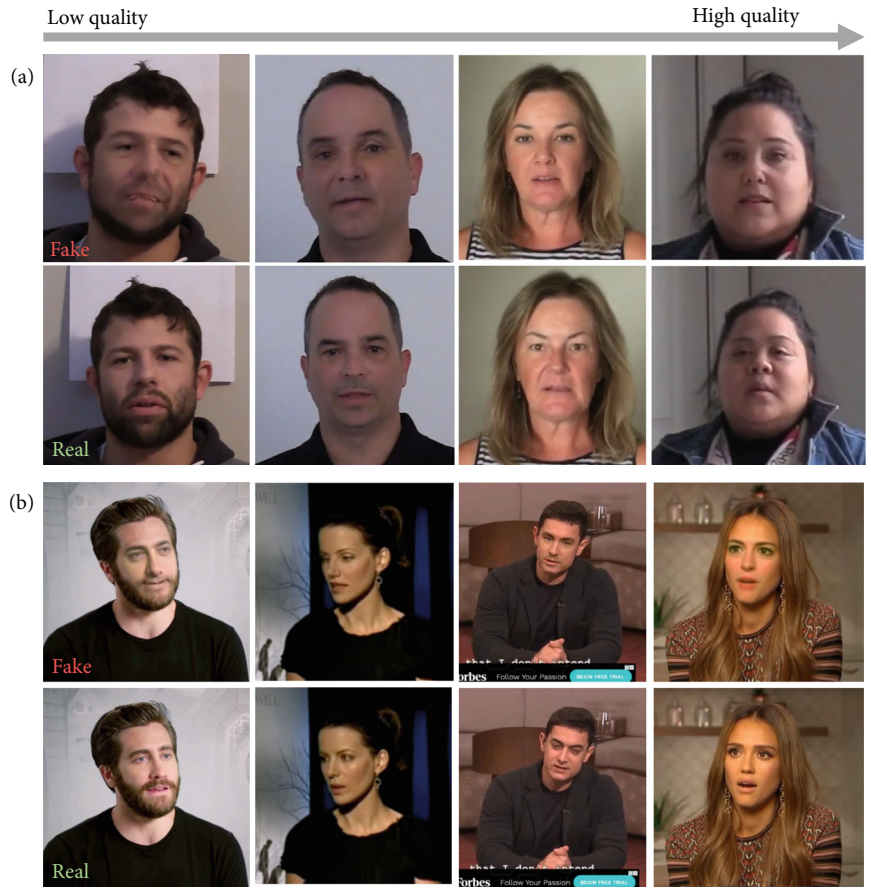

FIGURE 2: Different qualities of deepfakes between and within different datasets. *Note.* Images were taken from the (a) DeepFake Detection Challenge dataset and (b) Celeb-DF v2 dataset.

2.3. Data Extraction and Quality Assessment. The following information was extracted from the full-text articles by the first author: (1) details of the study (e.g., sample size, recruitment information); (2) details of the deepfake used (e.g., what type of media was used, what dataset was used); (3) intervention (if present) or variables assessed in relation to detection performance; (4) performance outcomes; and (5) how performance was measured.

The Joanna Briggs Institute (JBI) Critical Appraisal tools were used to assess the methodological rigor of the studies and to determine the extent to which a study has addressed and mitigated potential sources of bias in its design, conduct, and analysis. For the review, we specifically used the following checklists: (1) quasiexperimental studies [15], (2) analytical cross-sectional studies [16], and (3) qualitative research [17].

2.4. Synthesis Methods. We conducted a narrative synthesis of the data. Given the substantial variability in the design and outcomes measured in the included studies, we determined that a meta-analysis was not a suitable approach for synthesizing the findings.

3. Results

3.1. Study Selection. A flow diagram showing the selection process is provided in Figure 3. As shown, records were mostly excluded during the full-text screening stage for not using deepfake stimuli, using synthetic stimuli, lacking human participants, or not including a detection performance measure.

3.2. RQ1: What Existing Studies Have Examined People's Ability to Detect Deepfakes?

3.2.1. Study Characteristics. In total, 30 records with 40 independent studies were included at the data extraction stage. Sample sizes ranged from 19 to 11,088 and the majority of the records were published after 2020. Most were experimental studies ($n = 31$, 77.5% of the studies), and the media used as deepfake stimuli were mostly videos with audio ($n = 22$, 73.3% of the records). All audio stimuli were voice conversions. Almost all video/image stimuli used face swapping or lip-sync technique, except for three studies that used morphing techniques [6, 7, 18]. See Figure 4 for a visual breakdown of the characteristics of the reviewed studies (see Table S3 for the extended summary of the study characteristics).

Media were sourced from various datasets including DFDC [9], Perceptual Experiments on Face Swaps (PEFS) [19], trusted media challenge (TMC) [20], FaceForensics++ + [21], and Celeb-DF [10]. A small number of studies used stimuli that were created by the researchers themselves

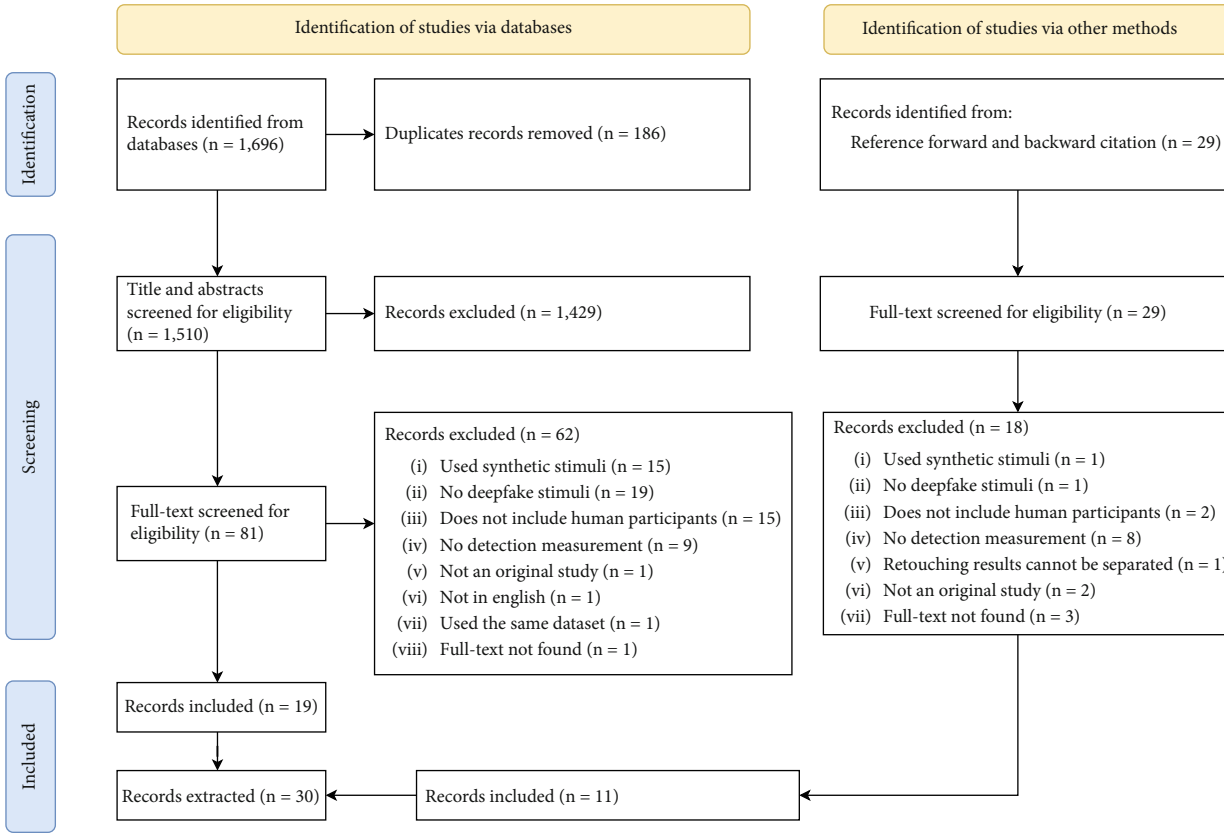

Figure 3: Flow diagram of the screening process.

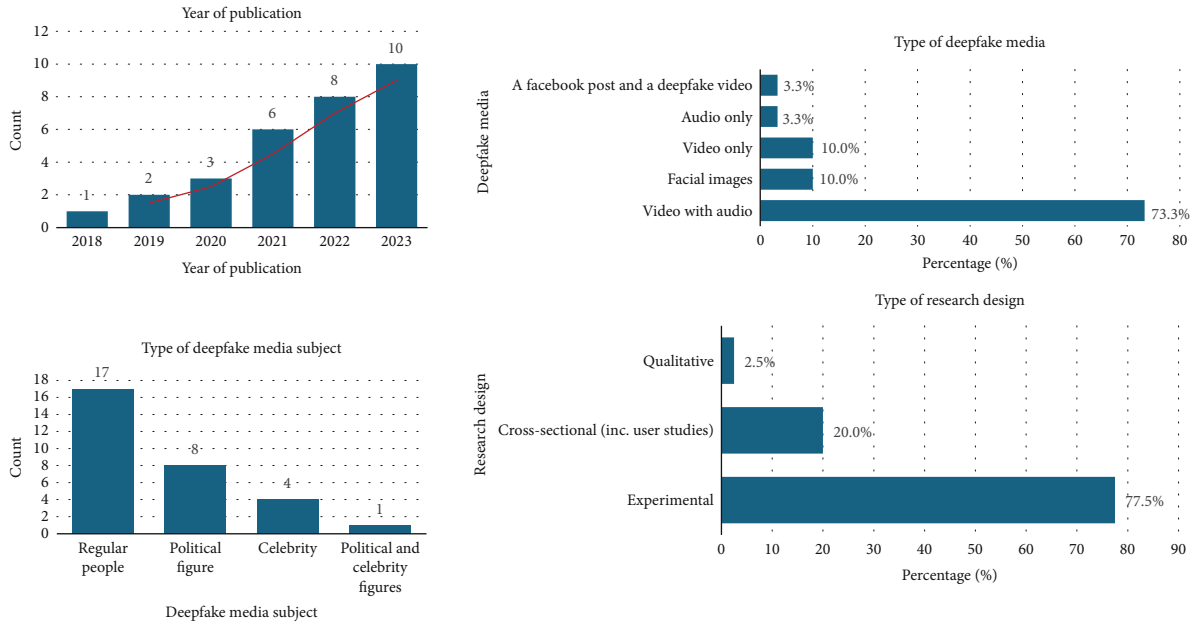

Figure 4: Characteristics of reviewed studies—year of publication, subject and type of deepfake media used, and type of research design.

($n = 7$, 23.3%) using software such as DeepFaceLab and JPsychmoph. One study asked participants to generate their own deepfakes [22]. Some studies also used popular and readily accessible deepfake videos of politicians created by organizations such as Future Advocacy, Buzzfeed, the Belgian Socialist Party, as well as from unknown sources. Most of the subjects in the deepfake videos were regular people (i.e., nonpopular figures; $n = 17$, 56.7%). Eight studies

used political deepfakes, four used popular celebrity deepfakes, and one study used both political and celebrity figures. The algorithms used for deepfake creation included deep learning techniques and generative adversarial networks (GANs). The stimuli presented to participants ranged from 1 deepfake clip to 100 deepfake–authentic clip pairings.

3.2.2. Risk of Bias Assessment. All records were assessed by the first author, and a random selection of 20% of full-text records ($n = 6$) was reviewed by another author (D.J.M.). In cases where there was a disagreement on a criterion, the two authors discussed the discrepancy until arriving at a collective decision (see Table S4 for the result of the risk of bias assessment). Most of the experiments were generally effective in addressing bias in their design, conduct, and analysis. However, none of the experimental studies had any follow-up, and only two studies had pre–post measurements of the outcome [6, 18]. Six of the experiments did not have a control group or control condition.

For the cross-sectional studies, the research design (or at the very least, reporting) could be improved. For example, only one study had clear inclusion criteria [23]. Only half of the studies clearly described their subjects and the research settings, and two studies did not clearly identify or control any potential confounding factors. One study was unclear about their analysis. There was only one qualitative study, and while there was congruence between the research methodology, questions, data analysis, and interpretation, other elements important to qualitative research, such as the researchers' philosophical perspective, localization, and participants' voice representation, were not reported. However, it is important to note that qualitative research in information sciences may not always follow the same reporting conventions as in the social sciences or health disciplines. In information sciences, qualitative studies often prioritize pragmatism over theoretical grounding or reflexive positionality of the researcher [24].

3.2.3. How Well Do People Perform in Deepfake Detection? As anticipated, performance varied significantly across studies depending on the deepfake dataset used as well as how detection was measured. As shown in Table 1, the studies reviewed used a variety of performance metrics to measure detection performance. These metrics include forced choices, Likert scales, open-ended questions, area under the curve (AUC), $F1$ score, precision, true positive rates, false positive rates, and false negative rates. These metrics are summarized in Table 1. In some studies, measures are converted into other metrics. For instance, Chen et al. [20] converted participant's Likert ratings into AUC to compare people's performance with AI detection models. In Groh et al.'s [2] study, participants' accuracy rating was calculated based on their correct identification minus confidence rating (i.e., if a participant responded "82% confident this is a DeepFake" and the participant was correct, then the participant was assigned an accuracy score of 0.82).

It is difficult to compare people's detection abilities across studies primarily due to the variety of metrics employed. The most common approach is presenting a stimulus to participants one at a time and directly asking participants to indicate whether the stimulus was fake or authentic ($n = 19$). Choices were often presented in either a two-forced choice format (yes, no), but some used a three-forced choice format (yes, no, I do not know). In some studies, participants were shown both the deepfake and real stimuli, at the same time, and participants had to choose which one was fabricated [2, 7].

With these direct measures, detection accuracy ranged from 57.6% to 75.43% (on average). These accuracy ratings were often then used to calculate other metrics such as AUC, $F1$ score, and recall. Metrics such as the AUC, $F1$ score, and recall are often used in machine learning to evaluate the performance of classification models, including those designed for deepfake detection, allowing human performance to be compared to machine learning detection models.

In some studies, judgment of the authenticity of videos was assessed on ordinal or continuous rating scales. Thirteen studies used a Likert scale format. For example, Dobber et al. [30] used the items, "I find the video authentic" and "I find the video credible" on a 1–7 Likert scale. In some cases, studies followed up Likert scales with open-ended questions asking participants to expand on their chosen rating. In Dobber et al.'s [30] study, participants who scored below 4 on either or both of the first two items were asked an open-ended question about why they considered the authenticity and credibility of the stimuli to be (somewhat) low.

Some studies measured participants' detection judgments in an indirect manner. For instance, Wohler et al. [35] asked participants "did you notice anything in the video?" to capture participants' perceptions while watching the videos. Detection was recorded if participants mentioned that the video was fake (or mentioned suspecting as much). In such studies, participants were not made aware that they were watching manipulated videos. The detection rates of these studies ranged from 0% to 83.5%.

Several studies assessed detection by asking participants to evaluate the veracity of specific statements made within the video (e.g., "did Obama called Trump a dipshit?" [26]), suggesting that deepfake detection may not solely depend on identifying visual or auditory manipulations but also scrutinizing the plausibility and truthfulness of the content presented. This metric is valuable when assessing deepfake videos of popular figures, particularly those in politics. Studies have shown the importance of assessing message content itself as people's agreement with the content of a message significantly affects judgments of the authenticity of videos [38, 39].

3.2.4. Humans Versus Computers. As mentioned previously, most studies on deepfake detection focused on the efficacy of detection algorithms. Within the studies reviewed, eight studies compared people's detection abilities with machine learning detection algorithms, but the algorithms used varied across studies. For instance, Groh et al. [2] used the machine learning detection model that won a competition hosted by Partnership for AI, Facebook, Microsoft, and Amazon. Out of 2116 submissions, the top model achieved a 65% accuracy rate on the DFDC dataset on the holdout set of 4000 videos (i.e., a portion

TABLE 1: Performance metrics used in the reviewed studies.

Metric	Performance measurement	Overall detection performance (and some examples)
Correct identification/ accuracy	Participants classify videos as fake or real, typically via forced-choice of yes/no or yes/no/I do not know. Some use side-by-side comparison	Participants' average score: 60.70% accuracy [4]; 57.6% [3]; 75.43% (forced-choice) and 69.92% (sequential) [7] Proportion of participants who correctly detected the fake videos: ~30% participants [25]; 50.8% not deceived, 16% were deceived, and 33.2% were uncertain [26]; 6/20 participants correctly identified all videos [27] Detection varies by video type and quality: Lip-sync: 52.6%, face swaps: 91.3%, real videos: 78.6% [28]; high quality: 66.57%, low quality: 58.73% [21]
Likert scale	Participants rated stimuli realism/authenticity/ credibility/technical manipulation on a Likert scale; sometimes assesses perceived accuracy of the claim made by the deepfake subject	Realism: Five options "real," "rather real," "rather fake," "fake," and "no idea." Responses of "no idea" (which constituted 14% of the total) were excluded from the analysis, and answers of "rather real" or "rather fake" were assigned a score of 0.5 [29]; authenticity/credibility: Likert scale of 1–7, deepfake video was rated 3.70 out 7, authentic video rated 4.18 out of 7 [30]; technically manipulated: Likert scale of 1–7, Boris Johnson's video was rated 5.34 vs. Barack Obama's video was rated 4.91 [23] Perceived accuracy of claim: Likert scale of 1–7 "how believable were the following elements of the political speech?" Average rating was 3.44 out of 7 [31]; Likert scale of 1–4, "…how accurate is the claim that Kim Kardashian said that she manipulates people online for money?" Average rating was 2.88 [32]; Likert scale of 1–7 of two items: (1) "this video is fake/real" and (2) "what Mr. Donald Trump/Mr. Barack Obama said in this video is fake/real." Donald Trump's video was rated 4.55 vs. Barack Obama's video was rated 3.19. Perceived message fakeness rating for Trump's video was 4.88 vs. Obama's video was 3.08 [33]
Open-ended questions	Detection measured by participants mentioning suspicion of inauthenticity	• Participants were interviewed and asked to identify which videos they watched were real or fake and what aspects of the videos that lead to their conclusions. 6 out of 20 participants correctly detected all videos [27] • When participants were asked to summarize the study's aim in their own words, none commented suspicion of video manipulation [34] • In a free description task, participants were asked to comment about video quality. In high-quality swaps, participants reported artifacts 44.44% of the videos vs. 83.49% for the lower quality swaps [35]
Area under the curve (AUC)	AUC represents the area under the ROC curve (which is a two-dimensional measure of classification performance—the true positive rate vs. false positive rate). Scores range from 0 to 1	• The best algorithm model: 0.985 vs. participants: 0.870 [20] • The best algorithm model: 0.7397 vs. participants 0.08747 [36] • The best algorithm model: 0.957 vs. participants: 0.936 [2]
*F*1 score and precision	*Precision* = true positives/true + false positives *F*1 *score* (aka *F-score*) combines precision and recall	*F*1 score = 54%, precision = 81% [37]
True positive rate (TPR)	*True positive rate* (TPR) aka *hit rate* or *recall* refers to the proportion of actual deepfake stimuli that were correctly identified as deepfake	TPR = 38% [37]
False positive rate (FPR) and false negative rate (FNR)	*False positive rate* (FPR) = authentic stimuli incorrectly classified as fake; *false negative rate* (FNR) = deepfake stimuli incorrectly classified as authentic	FPR = 23.8% – 42.6%; FNR = 24.0% – 41.0% [7]. Performance varied by manipulation difficulty

of the original dataset that is completely set aside and not used during model training or tuning). Groh et al. [2] compared human performance with the winning model using videos from DFDC and found that, among the participants who saw at least 10 videos, 82% outperformed the winning model.

In a second experiment, Groh et al. examined the *crowd mean*, which is the average response from participants for each video [2]. Using a sample of the holdout videos, the crowd mean accurately identified 74%–80% of the videos. The crowd's accuracy ratings were comparable to the winning model's accuracy rating of the sampled videos (which is 80% for the sampled videos). In another study that used DFDC videos, Korshonuv and Marcel [36] found that participants outperformed state-of-the-art machine learning detection models, achieving average AUC scores of 87.47% compared to EfficientNet-trained-on-Google at AUC of 73.97%.

Other studies have conflicting findings, with one study concluding that algorithms are "far better at detecting deepfake content than crowd cognitive abilities" [25]. Using the same DFDC dataset, Salini and HariKiran [25] found that approximately 70% of participants failed to detect the fake videos, with over 40% also misclassifying real videos as fake (i.e., false positives). In their study, detection algorithms achieved an accuracy rating of over 90% for the same videos. However, it is important to note that Salini and HariKiran [25] only presented participants with three pairs of videos. The authors also acknowledge that publicly available AI detection tools are currently nonexistent.

Using another deepfake dataset (TMC), Chen et al. [20] found that the best machine learning detection model achieved an AUC of 0.985. In comparison, the overall human performance scored an AUC of 0.870. In Kramer et al.'s [6] study of facial morphs, the automated detection model performed better than the best-performing human (68.4% vs. 64%, respectively), revealing a significant advantage for the automated detection model. Lastly, in the case of audio deepfakes, humans tend to perform very poorly compared to automated detection [29].

It is important to note the differences in performance between human and machine learning in deepfake detection. For instance, Korshonuv et al. [36] noted that machine learning algorithms struggled the most with deepfake videos that were easily discernible to human participants. Tahir et al. [37] suggested that this disparity could stem from differing focuses between humans and machines on detection tasks. Humans tend to prioritize main facial features such as eyes, hair, and nose, often overlooking smaller details, while detection systems prioritize more detailed aspects such as ear placement and facial outline. This attention disparity could be attributed to automated systems' capacity to analyze images down to individual pixels. Furthermore, Groh et al. [2] point out that the brain possesses a specialized region for processing faces holistically. In their research, humans were adversely affected by manipulations obstructing specialized face processing (e.g., inversion, misalignment, and eye occlusion), whereas computer performance remained unaffected. Overall, these studies suggest that people perform comparably to detection algorithms at detection tasks, possessing certain advantages that machines lack.

3.2.5. Eye-Tracking and Electroencephalography (EEG). Two studies employed eye-tracking methodology to detect people's performance in deepfake detection [35, 37]. In Wohler et al.'s [35] study, the distribution of eye fixations differed between real and high-quality face-swapped videos. For both real and low-quality videos, participants tended to fixate on the nose, mouth, and eyes, with no significant difference in their distributions. However, for high-quality face-swapped videos, participants tended to fixate on the nose and mouth, with fewer fixations on the eyes.

One study used EEG to assess cross-cultural perception of deepfakes [40]. Participants ($N = 10$) were more accurate in distinguishing between real and deepfake videos when the video actor was speaking in a language that the participant understood (e.g., English) and when the video actor was of a similar ethnic background to the participants. The EEG data suggested notable differences in brain activity patterns related to the perception of authentic and deepfake videos [40]. Brain activity patterns changing depending on whether people are watching real or deepfake stimuli highlights the adaptive nature of human perception, particularly in the context of digital media manipulation. The findings point to the potential for combining the strengths of both humans and AI in a collaborative effort—a human-in-the-loop (HITL) system. Such systems could leverage the unique abilities of both humans and machines, leading to more powerful and reliable deepfake detection systems [41].

3.3. RQ2: What Factors Influence People's Ability to Detect Deepfakes? The ability of people to detect deepfakes is influenced by a variety of factors, which can be categorized at the person level (factors inherent to individuals; see Table 2) or stimuli level (factors related to stimuli themselves; see Table 3). The effectiveness of potential interventions to improve deepfake detection is also outlined (see Table 4). For an overall summary of each study finding, see Table S5.

3.3.1. Person-Level Factors

3.3.1.1. Demographics. A few studies have examined the impact of demographic factors on one's ability to detect deepfakes, either as main predictors or control variables [32, 33, 37]. Overall, deepfake detection does not seem to vary based on gender, education, income, or profession [7, 32, 37]. Multiple studies found that ethnicity does not impact deepfake detection. However, one study found that deepfake detection improved when the detector was familiar with the language being spoken by the model or had a similar background to the model [40]. Doss et al. [43] observed that older age is associated with poorer detection. In Doss et al.'s [43] study, educators were more vulnerable to deepfakes than students. In contrast, other studies have found no relationship between detection and age [32, 33, 37].

3.3.1.2. Cognitive Ability, Analytical Thinking, and Truth Bias. Cognitive ability and analytical thinking seem to be associated with better detection of political and celebrity deepfakes. Appel and Prietzel [23] found that an individual's tendency to think analytically is correlated with general skepticism, which in turn is associated with being better able to discern between genuine and manipulated information. However, people also tend to display a "truth bias," a tendency to automatically assume that information is true

TABLE 2: Person-level factors impacting deepfake detection.

Factor	Influence on detection ability	Findings
Demographics		
Ethnicity	Mixed (—)	Ethnicity has no effect on detection ability [42] but people may be better at detecting deepfakes when the model is the same ethnicity as the detector or speaking a language the detector is familiar with [40].
Age	Mixed (—)	Mixed results found between age and detection accuracy: Doss et al. [43] found that older participants were more susceptible to deepfakes while others found that accuracy has no relationship with age [32, 33, 37].
Other demographics (gender, education, income, profession)	No effect (=/=)	Gender, education, income, and profession have no effects on detection performance [32, 42, 7, 37].
Cognitive ability/analytical thinking	Improves (↑)	Cognitive ability and analytical thinking are positively associated with detection accuracy [32, 23].
Truth bias (i.e., uncritically accepting belief of honesty)	Decreases (↓)	Truth bias is associated with miscategorizing deepfakes as real [44].
Political interest, knowledge, and distrust/partisanship	Mixed (—)	Political interest or partisanship has no association with deepfake detection performance across most studies [23, 32]. Those with conservative political attitudes may be less likely to correctly identify deepfakes, especially when video content aligns with their political views [38]. Political distrust and prior held beliefs about the video subject are associated with higher skepticism, regardless of whether the video is real or not [31].
Intuition and participant's emotional states	Mixed (—)	Individuals rely on intuition when doing detection tasks [27]. Anger is associated with being more likely to miscategorize real videos as fake [2].

TABLE 3: Stimuli-level factors impacting deepfake detection.

Factor	Influence on detection ability	Findings
Familiarity with the deepfaked subject	Improves (↑)	Familiarity with the video model was associated with better deepfake detection [29, 27].
Perceived character of the deepfaked subject	Decreases (↓)	Participants were more likely to categorize the real videos as fake when they viewed the depicted subjects (in this case, Trump and Obama) as dangerous. The opposite is observed for trust—participants who trusted the depicted politician were less likely to categorize the deepfake video as fake [33].
Quality of the deepfake video	Mixed (—)	Higher-quality deepfakes are associated with fewer visual artifacts, making them more difficult to detect [20, 35, 36, 42]. However, video perturbations can impact human judgment. Specifically, imperfections in authentic videos can lead to them being miscategorized as fake [2, 22, 37, 27, 35].
Type of stimuli (video vs. audio)	Mixed (—)	Mehta et al. [22] found that one-third of audio-only deepfakes were perceived as somewhat convincing but none of the combined audiovisual deepfakes were considered realistic enough to be believable. However, Ahmed and Chua [42] found no significant differences in perceived accuracy between video deepfakes and audio-only deepfakes.
Technique used for deepfake generation	Mixed (—)	Face swaps are easier to detect than morphs [7]. Lip-syncs are harder to detect than face-swapped videos [28]. Lip-syncs are easier to detect when there are unnatural or inconsistent lip/mouth movements [20, 30, 7, 28]. Morphs are easier to detect when they get to a 50% blend ratio. Error rates also differ depending on the quality of the morph [7].

by default [46]. In Son's [44] study, truth bias was associated with a higher likelihood of miscategorizing deepfake videos as real.

3.3.1.3. Political Interest, Knowledge, Distrust, and Partisanship. Some studies have assessed the impact of political interest, knowledge, distrust, and partisanship on detection of political

TABLE 4: Interventions to improve deepfake detection.

Factor	Influence on detection ability	Explanation
Raising awareness about deepfakes	No effect (=/=)	Raising awareness about the dangers of deepfakes has no impact on detection performance [3]
Warnings	Mixed (—)	Warnings that a video might be faked yields mixed results. Providing a warning label that a video may be deepfaked results in detectors doubting the authenticity of both deepfake and authentic videos [39]. Withholding warnings results in decreased likelihood of questioning the veracity of deepfake videos [32]
Revealing AI prediction	Improves (↑)	Providing the predictions of AI detection programs increases the performance of human detectors [2]
Financial incentives	No effect (=/=)	Financial incentives do not impact detection performance [3]
Media literacy training	Improves (↑)	Media literacy training improves detection performance [45]
Deepfake detection training with example videos	Improves (↑)	Deepfake detection training module with example videos improves detection accuracy [37]
Providing written tips	Mixed (—)	Mixed results as to whether providing written detection tips improves detection performance [6, 18, 4]

deepfakes. The effects of these variables are mixed. Overall, political interest and partisanship do not appear to have an impact on detection [23, 32]. However, one study found that participants with conservative political attitudes were less likely to correctly identify deepfakes, especially when the deepfake content aligned with their political views [38]. Hameeler et al. [31] also showed if a statement made by the video subject seemed out of character or unlikely for the subject to make (i.e., "*he would never say anything like this*"), participants were more likely to question the credibility of the video, regardless of whether the video was a deepfake or not. Hameeler et al.'s [31] study suggests that perceived authenticity is influenced by how well the content aligns with the participants' expectations of the video subject's character or typical behavior.

3.3.1.4. Intuition and Participants' Emotional State. Thaw et al. [27] found that participants rely on their intuition to determine whether videos are fake. Intuition, however, may be influenced by various factors, including emotional states. For instance, Groh et al. [2] induced anger in participants, and while anger did not affect overall detection accuracy, further analysis focusing solely on authentic videos revealed that angered participants were more prone to incorrectly identifying real videos as deepfakes. Thus, relying solely on intuition may lead to inaccurate assessments, particularly when emotions are heightened, as emotional responses appear to cloud judgment and critical analysis.

3.3.2. Stimuli-Level Factors. Given that deepfakes are computer-generated synthetic media, they often contain visual artifacts, such as poorly rendered eyes, odd skin textures, or imperfect lip-syncing. These artifacts can usually be detected with close examination. Thus, video features such as these can impact detection rates [2, 20, 22, 27, 35]. It is not surprising that higher-quality fakes (with fewer visual artifacts) are more difficult to detect than lower-quality fakes [20, 22, 35, 36]. However, some studies also showed that video imperfections (rather than computer-generated artifacts) can impair human judgment, leading detectors to miscategorize authentic videos as fake [2, 20, 22, 27].

Audio artifacts (e.g., distorted speech) can also occur. As mentioned above, humans tend to perform very poorly in detecting audio deepfakes [29], which may be attributable to the lack of more obvious visual cues. In a study by Mehta et al. [22], participants were instructed to assess the quality of audio and video deepfakes produced by inexperienced creators. When participants were asked to evaluate the audio clips in isolation (i.e., without the accompanying video), 34.76% of clips were reported as sounding "vaguely" like the person was being impersonated. However, when participants were presented with audio and video simultaneously, none of the stimuli were considered realistic enough to convince participants that the target individual was saying the intended statement. In contrast, Ahmed and Chua [42] did not find any significant differences in perceived accuracy levels between deepfake and audio deepfakes.

The technique used to generate deepfake stimuli can also impact detection performance. For instance, face swaps are easier to detect compared to facial morphs [7] and lip-sync videos are harder to detect than face swap videos [28]. However, unnatural lip movements, or speech and mouth movement inconsistencies, are critical tell-tale signs of manipulation [7, 20, 28, 30], suggesting that poor quality lip-syncs are quite easy to detect.

Detection performance also varies within the same deepfake generation technique. For instance, morphs are easier to detect when they get to a 50% ratio—i.e., equal blend of two faces [7]. Furthermore, Nichols et al. found that participants were more likely to make false-negative errors (i.e., saying a photo is fake, when it is real) for lower-quality morphs. The inverse was true for higher-quality morphs—participants were more likely to make false-positive errors (i.e., saying a photo is real, when it is fake). This suggests that the

relationship between error rates (FPR and FNR) reverses for easy and hard-to-detect morphs.

The identity of the person being deepfaked may also impact detection judgments. In general, familiarity with the deepfaked subject enhances one's ability to identify discrepancies or anomalies in the content [27, 29]. However, the perceived character of the depicted individual can complicate this effect. In Ng's [33] study, participants rated videos of Donald Trump as more fake (4.55 out of 7) than videos of Barack Obama (3.19 out of 7). When Trump or Obama was considered dangerous by participants, they were more likely to categorize the real videos as fake. The opposite is observed for trust—participants who trusted the depicted politician were less likely to categorize the deepfake video as fake [33].

3.3.3. Potential Interventions. A number of interventions to improve detection performance have been tested in the literature, showing mixed results. Raising awareness about the dangers of deepfakes or offering monetary incentives for correct detections do not impact detection performance [3]. Providing warnings that a video might be faked is associated with increased accuracy ratings [32]. However, providing warning labels may have the unintended consequence of causing participants to doubt the authenticity of all videos watched, regardless of whether they fake or not [39]. Other studies investigated the effectiveness of targeted detection strategies in the form of written tips outlining common visual artefacts found in deepfakes. Some examples of these tips include instructions to pay attention to whether "skin appears too smooth or too wrinkly" [4] or smoother than in normal photographs [6] or to look for "ghost-like" outlines around the face [18]. Providing this kind of basic detection guidance yielded no improvement on detection accuracy [4, 6]. In contrast, Robertson et al.'s [18] study involved an additional element of training whereby feedback was provided to participants after each trial, which resulted in improved detection performance.

Intervention that incorporated interactive training seemed to be the most effective, especially if coupled with example videos and feedback. Studies that utilized a similar approach include El Mokadem's [45] study whereby participants went through a media literacy training program based on Inoculation Theory delivered via a lecture format. The lectures highlighted the threat of misinformation and deepfakes. Examples of misinformation posts and deepfake videos were also provided, with guidance on how to detect them and how to verify content on social media. Tahir et al. [37] developed a short training module with example videos highlighting visual anomalies (e.g., skin discoloration, facial flickering, and unnatural features). Trainees were warned against relying solely on inconclusive factors, like unprofessional looking logos. In both studies, participants who underwent the training showed better deepfake detection performance compared to a control group.

Another promising intervention is to provide humans with predictions generated by AI-based detection models. Groh et al. [2] found that participants performed better on a detection task when shown an AI's assessment of a video's authenticity. This suggests that AI can serve as a valuable tool in assisting humans to identify deepfakes more effectively.

4. Discussion

This review was aimed at synthesizing the current state of research on human deepfake detection. In total, 40 studies from 30 records were reviewed. This review differs from previous reviews, which focus primarily on the performance of automated deepfake detectors [1]. Despite the general efficacy of automated deepfake detectors, this technology is yet to be deployed on a mass scale (e.g., integrated into social media feeds), meaning that humans are still largely reliant on their own detection abilities when exposed to deepfakes while browsing social media. Thus, it is crucial to understand human performance in detecting AI-manipulated media.

4.1. Person-Level and Stimuli-Level Factors Impacting Deepfake Detection. Our review found that detection accuracy varied across studies, highlighting the complexity and challenges of identifying manipulated media. This variation in performance can be attributed to methodological variations between studies and also to a range of other factors occurring at both the person and stimuli level. At the person level, higher cognitive ability, analytical thinking skills, and the tendency not to accept information at face value are all associated with better deepfake detection [23, 32]. Demographic factors such as gender, education, and income level appear to have no discernible effect on detection performance [7, 32, 37, 42], while studies into the effects of ethnicity and age yield inconsistent results [32, 37, 43]. Interest in and knowledge of politics and political distrust demonstrate inconsistent effects on detection performance [23, 32], so too do intuition and emotional arousal [2, 27].

At the stimuli level, greater familiarity with the deepfaked subject and poorer quality deepfakes detect are both associated with better deepfake detection. However, several caveats should be discussed. First, the relationship between familiarity and detection accuracy is complicated by viewers' pre-existing perceptions of the depicted individuals [33]. Second, while lower-quality deepfakes are generally easier to detect [20, 35, 36, 42], imperfections in authentic videos (e.g., lighting inconsistencies or unusual framing) can lead viewers to miscategorize genuine content as manipulated [2, 22, 27, 35, 37]. This suggests that detection accuracy depends not only on the presence of deepfake-specific artifacts but also on viewers' ability to distinguish between intentional manipulation and incidental quality issues in digital media. Type of stimuli (video versus audio) and the technique used for deepfake generation (e.g., face swapping, lip-syncing, and level of morphing) demonstrated mixed results [7, 20, 22, 28, 30, 42]. This variability likely stems from the differing perceptual demands each manipulation type places on viewers.

4.2. Consistency in Performance Metrics. We also found that it was difficult to compare people's detection abilities across studies primarily due to the variety of metrics employed to measure detection. Given the variability in measurement found in the studies reviewed, we provide the following

recommendations to standardize the field of human detection of deepfakes, with an overall aim to improve reporting and ease of quantifying effects across this literature. First, we suggest that authors avoid using Likert scales or other continuous measures in relation to the categorization of stimuli as fake or real, as these kinds of nondichotomous response formats impede the calculation of useful statistics such as hit rate and specificity.

If authors do wish to make use of nondichotomous response formats (for example, to capture participant uncertainty), they may consider either a three-point format (fake, unsure, real) or the use of a Likert-type scale with *real* and *fake* as anchor points (see Figure 5). If using Likert-type scales with even-numbered response options, data can be dichotomized during analysis (e.g., grouping response options 1–3 and 4–6 as categorizations of fake and real, respectively), although issues regarding dichotomization of continuous variables should be noted (see [47]).

When using dichotomous response formats, overall accuracy can be calculated as a simple and intuitive measure of performance. The mean accuracy of a sample may be reported in a percentage (e.g., "on average, the control group correctly categorized 61.4% of stimulus videos") or raw score format. In the latter case, the number of stimuli presented to participants should be made clear so that raw scores can be converted to percentages if needed ("of the 15 stimulus videos presented, participants correctly categorized 9 on average"). It is also recommended that researchers indicate the accuracy that would be expected by chance alone to provide context for accuracy scores.

We recommend that researchers, in addition to reporting overall accuracy, also report participant performance statistics separately for real and deepfake stimuli. This information can be presented in a confusion matrix (which reports the frequency of true positives, true negatives, false positives, and false negatives; see Figure 6). From the confusion matrix, several elementary classification measures can be derived, such as accuracy, error rate, sensitivity, specificity, precision and the *F* measure. Some of these (e.g., *F* measure) are frequently reported in relation to the performance of machine learning models. Providing the information required to calculate these measures would therefore facilitate the comparison of human raters under various conditions to machine learning models. For studies that draw randomly from video databases (e.g., [20]), accuracy can be presented at the trial level (e.g., "2702 out of 4510 fake trials were correctly identified").

Researchers should also report the characteristics of the stimuli used, at the bare minimum, the source of the stimuli, the type of models depicted (whether celebrities, political figures, regular people, or a mixture), and information about the intervention if applicable. These factors can influence detection difficulty; thus, reporting these kinds of details can help contextualize findings.

4.3. Interventions to Improve Human Detection of Deepfakes. The ease with which high-quality deepfakes can now be generated, coupled with the potential for this technology to be used maliciously, has resulted in research interest in developing interventions to enhance the public's ability to detect deepfakes. Despite this growing interest, it appears that, with a few exceptions, the interventions applied up to this point have been ineffective. Incentivizing performance, by raising awareness of the dangers of deepfakes or providing financial incentives for good performance, does not appear to bolster detection [3], while providing warnings about the potential for a video being deepfaked seems to increase detection of deepfakes, but at the expense of mistrust of authentic videos [39]. Furthermore, explicitly teaching detection skills by providing written detection tips does not appear to improve detection compared to control [4, 6].

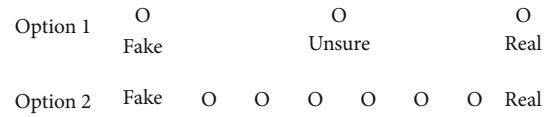

Figure 5: Suggestions for Likert-type scales.

These findings suggest that interventions should go beyond raising awareness about deepfakes and the use of labeling techniques (as in recent social media initiatives to label AI-generated content; Bickert, [48]). Our review suggests that media literacy training may be effective in improving performance. Interactive training activities in which feedback and/or example videos are provided also appear to bolster detection performance [18, 37]. Therefore, the most promising approach to improve detection ability, at least in the short term, appears to be general media literacy training focused on developing analytical thinking skills, coupled with interactive training activities in which immediate feedback is provided. Additional testing may be required to assess whether incorporating practice sessions with increasing difficulty levels can further build detection performance.

4.4. Human–AI Collaboration. As noted, humans and AI systems tend to focus on different aspects during deepfake detection. On the one hand, AI can quickly analyze vast amounts of data at a granular level, focusing on specific details that may elude the human eye. For example, FakeCatcher detects synthetic videos by detecting changes in faces to infer blood flow [49]. FakeCatcher assesses whether skin color changes naturally over time as the heart pumps blood and whether there is coherence across facial regions. In one test, the detector reached a 91% accuracy rating.

Humans, on the other hand, excel at holistic facial processing and can process contextual information. For instance, an AI-generated image of Joe Biden was circulated online, depicting Biden in military fatigues while seemingly in discussion with military personnel [50]. However, some people may be aware that it would be atypical for a president to dress in a military combat uniform and thus may be skeptical of the image based on their prior contextual understanding.

Consideration should be given to how to integrate feedback from AI-based deepfake detection models to bolster human performance (and vice versa). Given the speed at which AI can process videos, there is scope for AI to be used to do an initial "first-pass" check for manipulated media.

		Reality		
		Deepfake	Real	
Decision	Deepfake	True positive (hit)	False positive (false alarm)	TP + FP
	Real	False negative (miss)	True negative (specificity)	FN + TN
		TP + FN	FP + TN	TP + FP + FN + TN

Figure 6: Confusion matrix.

Knowing that there will be errors, human raters can then focus on classifying these errors. Feedback from humans can then be used to refine the AI algorithms, making algorithms more attuned to the types of errors that humans are good at catching. AI systems could be trained to mimic human scanning patterns of faces, potentially improving their ability to detect deepfakes in ways that humans find intuitive. This could involve training algorithms on datasets that have been annotated according to human focus areas like eyes, nose, and mouth, as suggested by the differing focuses observed between humans and machines [35]. Furthermore, humans are capable of providing context and analyzing and tracing sources of information, processes that are still hard to automate [51].

The complementary nature of human and AI capabilities suggests that there is potential to leverage human–AI collaboration to improve deepfake detection. By developing systems that combine the strengths of both humans and machines, we may be able to achieve higher levels of detection accuracy and resilience against increasingly sophisticated manipulations.

4.5. Limitations and Future Studies. The current review focused on deepfakes in which the likeness of an existing person was imitated. Accordingly, studies that examined face retouching or face or voice synthesis were excluded from this review. This narrower focus limits the generalizability of the findings to these types of manipulated digital media.

Furthermore, it is also important to recognize that deepfake technology is continuously evolving and that the rapid development of these technologies may limit the applicability of the findings to the latest deepfake techniques. For instance, there is currently very little research on audio-based deepfakes. As new deepfake methods emerge and existing ones become more sophisticated, the strategies and findings discussed in this review will need to be updated and reassessed.

This review was limited to English-language publications. We acknowledge that because of this, our review is restricted in its global representation and thus may have overlooked important cultural and societal factors that could influence detection accuracy. For example, countries are likely to differ on sociocultural variables, like trust in government and media, variables which could impact general skepticism toward videos. Future research should investigate these possibilities.

Another limitation to consider is the lack of longitudinal data in the included studies. Most of the studies in this review are cross-sectional or experimental. This makes it difficult to draw conclusions about the long-term effectiveness of deepfake detection interventions or how detection abilities are changing as deepfakes become more embedded in our popular media over time. We suggest interventions that can be consistently demonstrated to be efficacious in the short term should be longitudinally investigated to demonstrate that these training effects are not merely transitory.

4.6. Conclusion. In conclusion, this systematic review provides valuable insights into the current state of research on human deepfake detection. The findings highlight the complex interplay of factors influencing people's ability to identify manipulated media, including person-level characteristics, stimuli-level features, and potential interventions. While some studies suggest that humans perform comparably to, or even outperform, machine learning algorithms in certain contexts, the rapidly evolving nature of deepfake technology presents ongoing challenges. The general lack of efficacy for the tested interventions underscores the need for further research to inform our understanding of how to enhance human detection capabilities. As deepfakes become increasingly sophisticated and prevalent, a multifaceted approach that combines technological tools, human expertise, and media literacy education may be necessary to combat the threats posed by AI-manipulated media. Collaboration between humans and AI, leveraging their respective strengths, is a promising solution. By shedding light on the current landscape of human deepfake detection research, this review provides a foundation for future studies and informs the development of countermeasures to mitigate the potential harms associated with deepfakes.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Conflicts of Interest

The authors declare no conflicts of interest.

Funding

No funding was received for this manuscript.

Supporting Information

Additional supporting information can be found online in the Supporting Information section. This manuscript is accompanied by supporting information providing additional information to support the main findings presented. The supporting information includes the following.
Supporting Information. Table S1: PRISMA checklists. Table S2: Search strategy used for each database. Table S3: Extended summary of the study characteristics detailing the author, year, research design, sample characteristics, deepfake stimuli, type of deepfaked figure (political, celebrity, or regular people), intervention details, and detection measurement. Table S4: Risk of bias assessment. Table S5: Overall summary of each study findings.

References

[1] R. Tolosana, R. Vera-Rodriguez, J. Fierrez, A. Morales, and J. Ortega-Garcia, "Deepfakes and Beyond: A Survey of Face Manipulation and Fake Detection," *Information Fusion* 64 (2020): 131–148, https://doi.org/10.1016/j.inffus.2020.06.014.

[2] M. Groh, Z. Epstein, C. Firestone, and R. Picard, "Deepfake Detection by Human Crowds, Machines, and Machine-Informed Crowds," *Proceedings of the National Academy of Sciences* 119, no. 1 (2022): e2110013119, https://doi.org/10.1073/pnas.2110013119.

[3] N. C. Kobis, B. Dolezalova, and I. Soraperra, "Fooled Twice: People Cannot Detect Deepfakes but Think They Can," *iScience* 24, no. 11 (2021): https://doi.org/10.1016/j.isci.2021.103364.

[4] K. Somoray and D. J. Miller, "Providing Detection Strategies to Improve Human Detection of Deepfakes: An Experimental Study," *Computers in Human Behavior* 149 (2023): 107917, https://doi.org/10.1016/j.chb.2023.107917.

[5] C. Bregler, M. Covell, and M. Slaney, "Video Rewrite: Driving Visual Speech With Audio," in *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, ed. M. C. Whitton (ACM, 1997), 715–722, https://doi.org/10.1145/3596711.3596787.

[6] R. S. Kramer, M. O. Mireku, T. R. Flack, and K. L. Ritchie, "Face Morphing Attacks: Investigating Detection With Humans and Computers," *Cognitive Research: Principles and Implications* 4, no. 1 (2019): https://doi.org/10.1186/s41235-019-0181-4.

[7] R. Nichols, C. Rathgeb, P. Drozdowski, and C. Busch, "Psychophysical Evaluation of Human Performance in Detecting Digital Face Image Manipulations," *IEEE Access* 10 (2022): 31359–31376, https://doi.org/10.1109/ACCESS.2022.3160596.

[8] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfakes Generation and Detection: State-of-the-Art, Open Challenges, Countermeasures, and Way Forward," *Applied Intelligence* 53, no. 4 (2023): 3974–4026, https://doi.org/10.1007/s10489-022-03766-z.

[9] B. Dolhansky, J. Bitton, B. Pflaum, et al., "The DeepFake Detection Challenge (DFDC) Dataset" 2020, https://arxiv.org/abs/2006.07397.

[10] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu, "Celeb-DF: A Large-Scale Challenging Dataset for DeepFake Forensics" 2020, https://arxiv.org/abs/1909.12962.

[11] L. A. Passos, D. Jodas, K. A. da Costa, L. A. S. Júnior, D. Colombo, and J. P. Papa, "A Review of Deep Learning-Based Approaches for Deepfake Content Detection" 2022, https://arxiv.org/abs/2202.06095.

[12] D. Fallis, "The Epistemic Threat of Deepfakes," *Philosophy & Technology* 34, no. 4 (2021): 623–643, https://doi.org/10.1007/s13347-020-00419-2.

[13] M. J. Page, J. E. McKenzie, P. M. Bossuyt, et al., "The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews," *BMJ* 372 (2021): https://doi.org/10.1136/bmj.n71.

[14] J. Clark, P. Glasziou, C. D. Mar, A. Bannach-Brown, P. Stehlik, and A. M. Scott, "A Full Systematic Review Was Completed in 2 Weeks Using Automation Tools: A Case Study," *Journal of Clinical Epidemiology* 121 (2020): 81–90, https://doi.org/10.1016/j.jclinepi.2020.01.008.

[15] T. H. Barker, N. Habibi, E. Aromataris, et al., "The Revised JBI Critical Appraisal Tool for the Assessment of Risk of Bias for Quasi-Experimental Studies," *JBI Evidence Synthesis* 22, no. 3 (2024): 378–388, https://doi.org/10.11124/JBIES-23-00268.

[16] S. Moola, Z. Munn, C. Tufanaru, et al., "Chapter 7: Systematic Reviews of Etiology and Risk," in *JBI Manual for Evidence Synthesis*, eds. E. Aromataris and Z. Munn (JBI, 2020), https://synthesismanual.jbi.global.

[17] C. Lockwood, Z. Munn, and K. Porritt, "Qualitative Research Synthesis: Methodological Guidance for Systematic Reviewers Utilizing Meta-Aggregation," *International Journal of Evidence-Based Healthcare* 13, no. 3 (2015): 179–187, https://doi.org/10.1097/XEB.0000000000000062.

[18] D. J. Robertson, A. Mungall, D. G. Watson, K. A. Wade, S. J. Nightingale, and S. Butler, "Detecting Morphed Passport Photos: A Training and Individual Differences Approach," *Cognitive Research: Principles and Implications* 3 (2018): 1–11, https://doi.org/10.1186/s41235-018-0113-8.

[19] L. Wöhler, J.-O. Henningson, S. Castillo, et al., "PEFS: A Validated Dataset for Perceptual Experiments on Face Swap Portrait Videos," in *Computer animation and social agents*, eds. F. Tian, X. Yang, D. Thalmann, W. Xu, J. J. Zhang, N. M. Thalmann, and J. Chang (Springer International Publishing, 2020), 120–127, https://doi.org/10.1007/978-3-030-63426-1_13.

[20] W. Chen, S. L. B. Chua, S. Winkler, and S.-K. Ng, "Trusted Media Challenge Dataset and User Study," in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM '22)* (Association for Computing Machinery, 2022), 3873–3877, https://doi.org/10.1145/3511808.3557715.

[21] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner, "Faceforensics++: Learning to Detect Manipulated Facial Images," in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (IEEE/CVF, 2019), https://openaccess.thecvf.com/content_ICCV_2019/html/Rossler_FaceForensics_Learning_to_Detect_Manipulated_Facial_Images_ICCV_2019_paper.html.

[22] P. Mehta, G. Jagatap, K. Gallagher, et al., "Can Deepfakes Be Created on a Whim?," in *Companion Proceedings of the ACM Web Conference 2023* (ACM, 2023), 1324–1334, https://doi.org/10.1145/3543873.3587581.

[23] M. Appel and F. Prietzel, "The Detection of Political Deepfakes," *Journal of Computer-Mediated Communication* 27, no. 4 (2022): 13, https://doi.org/10.1093/jcmc/zmac008.

[24] G. Goldkuhl, "Pragmatism vs Interpretivism in Qualitative Information Systems Research," *European Journal of Information Systems* 21, no. 2 (2012): 135–146, https://doi.org/10.1057/ejis.2011.54.

[25] Y. Salini and J. HariKiran, "DeepFake Videos Detection Using Crowd Computing," *International Journal of Information Technology* 16, no. 7 (2024): 4547–4564, https://doi.org/10.1007/s41870-023-01494-2.

[26] C. Vaccari and A. Chadwick, "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News," *Social Media + Society* 6, no. 1 (2020): https://doi.org/10.1177/2056305120903408.

[27] N. N. Thaw, T. July, A. N. Wai, D. H. L. Goh, and A. Y. K. Chua, "Is it Real? A Study on Detecting Deepfake Videos," *Proceedings of the Association for Information Science and Technology* 57, no. 1 (2020): https://doi.org/10.1002/pra2.366.

[28] I. Sundström, "Deepfake Detection by Humans: Face Swap Versus Lip Sync" 2023, https://kth.diva-portal.org/smash/get/diva2:1801193/FULLTEXT01.pdf.

[29] V. Barnekow, D. Binder, N. Kromrey, P. Munaretto, A. Schaad, and F. Schmieder, "Creation and Detection of German Voice Deepfakes," in *Lecture Notes in Computer Science*, eds. E. Aimeur, M. Laurent, R. Yaich, B. Dupont, and J. Garcia-Alfaro (13291, LNCS. Springer Science and Business Media Deutschland GmbH; Scopus, 2022), https://doi.org/10.1007/978-3-031-08147-7_24.

[30] T. Dobber, N. Metoui, D. Trilling, N. Helberger, and C. De Vreese, "Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes?," *International Journal of Press/Politics* 26, no. 1 (2021): 69–91, https://doi.org/10.1177/1940161220944364.

[31] M. Hameleers, T. G. van der Meer, and T. Dobber, "They Would Never Say Anything Like This! Reasons to Doubt Political Deepfakes," *European Journal of Communication* 39, no. 1 (2024): 56–70, https://doi.org/10.1177/02673231231184703.

[32] S. Ahmed, "Fooled by the Fakes: Cognitive Differences in Perceived Claim Accuracy and Sharing Intention of Non-Political Deepfakes," *Personality and Individual Differences* 182 (2021): 111074, https://doi.org/10.1016/j.paid.2021.111074.

[33] Y. L. Ng, "An Error Management Approach to Perceived Fakeness of Deepfakes: The Moderating Role of Perceived Deepfake Targeted Politicians' Personality Characteristics," *Current Psychology* 42, no. 29 (2023): 25658–25669, https://doi.org/10.1007/s12144-022-03621-x.

[34] A. Eberl, J. Kühn, and T. Wolbring, "Using Deepfakes for Experiments in the Social Sciences-a Pilot Study," *Frontiers in Sociology* 7 (2022): 907199, https://doi.org/10.3389/fsoc.2022.907199.

[35] L. Wöhler, M. Zembaty, S. Castillo, and M. Magnor, "Towards Understanding Perceptual Differences Between Genuine and Face-Swapped Videos," in *Proceedings of the 2021 CHI conference on human factors in computing systems* (ACM, 2021), https://doi.org/10.1145/3411764.3445627.

[36] P. Korshunov and S. Marcel, "Deepfake Detection: Humans vs. Machines" 2020, https://arxiv.org/abs/2009.03155.

[37] R. Tahir, B. Batool, H. Jamshed, et al., "Seeing Is Believing: Exploring Perceptual Differences in Deepfake Videos," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (ACM, 2021), https://doi.org/10.1145/3411764.34456.

[38] S. Sütterlin, T. F. Ask, S. Mägerle, et al., "Individual Deep Fake Recognition Skills Are Affected by Viewer's Political Orientation, Agreement With Content and Device Used," in *Augmented Cognition*, eds. D. D. Schmorrow and C. M. Fidopiastis (Springer Nature Switzerland, 2023), 269–284, https://doi.org/10.1007/978-3-031-35017-7_18.

[39] J. Ternovski, J. Kalla, and P. M. Aronow, "Deepfake Warnings for Political Videos Increase Disbelief but Do Not Improve Discernment: Evidence From Two Experiments" 2021, https://osf.io/dta97_v1.

[40] M. R. Khan, S. Naeem, U. Tariq, et al., "Exploring Neurophysiological Responses to Cross-Cultural Deepfake Videos," in *Proceedings of the Companion Publication of the 25th International Conference on Multimodal Interaction* (ACM, 2023), https://doi.org/10.1145/3610661.3617148.

[41] L. Jiang, R. Li, W. Wu, C. Qian, and C. C. Loy, "DeeperForensics-1.0: A Large-Scale Dataset for Real-World Face Forgery Detection" 2020, https://arxiv.org/abs/2001.03024.

[42] S. Ahmed and H. W. Chua, "Perception and Deception: Exploring Individual Responses to Deepfakes Across Different Modalities," *Heliyon* 9, no. 10 (2023)https://www.cell.com/heliyon/fulltext/S2405-8440(23)07591-6.

[43] C. Doss, J. Mondschein, D. Shu, T. Wolfson, D. Kopecky, V. A. Fitton-Kane, et al., "Deepfakes and Scientific Knowledge Dissemination," *Scientific Reports* 13, no. 1 (2023): 13429, https://doi.org/10.1038/s41598-023-39944-3.

[44] K. Son, "Deepfakes in Digital Discourse: The Impact of Information Priming and Truth Bias on Deception Detection" 2022, https://sites.lsa.umich.edu/commmediathesisprogram/wp-content/uploads/sites/1393/2025/01/SON-2022.pdf.

[45] S. S. El Mokadem, "The Effect of Media Literacy on Misinformation and Deep Fake Video Detection," *Arab Media & Society* 35, no. 35 (2023): https://doi.org/10.70090/SM23EMLM.

[46] T. R. Levine, "Truth-Default Theory (TDT): A Theory of Human Deception and Deception Detection," *Journal of Language and Social Psychology* 33, no. 4 (2014): 378–392, https://doi.org/10.1177/0261927X14535916.

[47] R. C. MacCallum, S. Zhang, K. J. Preacher, and D. D. Rucker, "On the Practice of Dichotomization of Quantitative Variables," *Psychological Methods* 7, no. 1 (2002): 19, https://doi.org/10.1037/1082-989x.7.1.19.

[48] M. Bickert, "Our approach to labeling AI-generated content and manipulated media" Meta, Apr. 5, 2024. [Online]. Available: https://about.fb.com/news/2024/04/metas-approach-to-labeling-ai-generated-content-and-manipulated-media/.

[49] U. A. Ciftci, I. Demir, and L. Yin, "FakeCatcher: Detection of Synthetic Portrait Videos Using Biological Signals," in *Proceedings of the IEEE Transactions on Pattern Analysis and Machine Intelligence* (IEEE, 2020), https://doi.org/10.1109/TPAMI.2020.3009287.

[50] S. A. Thompson, "A.I. Is Getting Better Fast," *Can You Tell What's Real Now?* (2024, The New York Times. https://www.nytimes.com/interactive/2024/06/24/technology/ai-deepfake-facebook-midjourney-quiz.html.

[51] N. Capuano, G. Fenza, V. Loia, and F. D. Nota, "Content-Based Fake News Detection With Machine and Deep Learning: A Systematic Review," *Neurocomputing* 530 (2023): 91–103, https://doi.org/10.1016/j.neucom.2023.02.005.