Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

AI assistance and error communication in deepfake evidence evaluation: A systematic review of human authenticity judgments

Hardianto Djanggih

hardianto.djanggih@umi.ac.id

Universitas Muslim Indonesia https://orcid.org/0009-0004-3911-2357

Systematic Review

Keywords: AI assistance · Criminal evidence · Deepfake detection · Error-rate disclosure · Human–AI collaboration · Systematic review

Posted Date: October 23rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11460998/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Abstract

Objectives

Assess how case-specific AI assistance and error communication affect human judgments of visual deepfakes and identify implications for criminal evidence.

Methods

Europe PMC, OpenAlex, and targeted citation follow-up were searched through 20 October 2026. A single AI-assisted workflow screened 182 unique records. Comparative human studies were synthesized without meta-analysis.

Results

Seven reports were included. Benefits depended on advice quality and presentation; incorrect guidance could worsen decisions. Disclosed reliability and actual performance were not consistently aligned. Artifact amplification improved detection in selected settings but did not establish robustness to erroneous assessments. No included report tested criminal verdict accuracy.

Conclusions

Assistance, trust, and accurate judgment are distinct outcomes. Forensic communication should state validation conditions and inferential limits. Restricted index coverage, heterogeneous designs, and absence of independent duplicate screening limit certainty.

Introduction

Digital recordings can influence criminal investigations because they appear to preserve what a person said or did. Deepfake generation complicates that inference by making facial appearance, movement, and speech reproducible without the represented event occurring. A further problem arises when an evaluator encounters a detector’s opinion about the recording. The evaluator must assess not only the media but also whether the accompanying technical assessment deserves reliance. This review examines that second problem: how people change authenticity judgments when given case-specific AI assistance and information about its performance.

Deepfake is an umbrella term rather than a single forensic condition. Reviews distinguish face synthesis, identity substitution, attribute editing, and expression manipulation, with different generation processes and detectable traces (Verdoliva 2020; Mirsky and Lee 2021; Tolosana et al. 2020). The associated harms also extend beyond deception to reputational injury, impersonation, and loss of confidence in

authentic recordings (Westerlund 2019; Chesney and Citron 2019). An experiment on one manipulation family therefore supplies evidence about a bounded task, not the authenticity of all digital media.

The technical and behavioral literatures address different units of analysis. FaceForensics + + and the DeepFake Detection Challenge supply datasets and benchmarks for evaluating detectors across specified manipulations and processing conditions (Rössler et al. 2019; Dolhansky et al. 2020). They do not, by themselves, establish how an investigator interprets a score or how a judge weighs the resulting opinion. A useful review must connect these literatures without treating machine classification and legally justified inference as interchangeable outcomes.

The relevant outcome is not simply greater confidence in detection technology. An intervention can increase agreement with a detector while causing additional errors on the cases the detector gets wrong. Conversely, declining to follow advice can preserve a correct independent judgment or discard a useful correction. The broader distinction between trust and appropriate reliance is established in human–automation research (Lee and See 2004). For evidentiary purposes, assistance should be evaluated by the quality of resulting judgments and their sensitivity to the advice's correctness, not by satisfaction or adoption alone.

Existing reviews provide useful accounts of overall human deepfake detection and interventions across different modalities (Diel et al. 2024; Somoray et al. 2025). Their broad questions do not eliminate the need to examine the narrower interaction between a person and an AI assessment. Training, generic warnings, independent machine benchmarking, and advice delivered during a decision are different interventions. Evidence that a detector outperforms unassisted observers does not establish that observers will benefit when shown its output. Evidence that explanations are liked does not establish that they improve discrimination.

Human vulnerability is also conditional. Research on synthetic faces, overconfidence in video judgments, political-speech modalities, and multimodal synthetic content demonstrates why performance must be tied to the stimulus and task (Nightingale and Farid 2022; Köbis et al. 2021; Groh et al. 2024; Cooke et al. 2025). These contextual studies do not estimate the intervention effect examined here. Their importance is that confidence, perceived trustworthiness, and discrimination can diverge. Uncertainty may also undermine trust in news even when audiences are not straightforwardly deceived (Vaccari and Chadwick 2020). For criminal evidence, both misplaced belief and unwarranted rejection of authentic material deserve examination.

The criminological question concerns the interpretation of technically mediated evidence. A deepfake classifier may address whether a file contains specified synthetic features. It does not, by itself, establish who produced or transmitted the file, whether the depicted conduct occurred, or whether an accused person committed an offence. Misunderstanding an authenticity assessment can therefore affect both acceptance of fabricated material and dismissal of genuine material. These are plausible institutional consequences motivating the review; the included studies do not directly measure wrongful convictions or case outcomes.

The review asks three questions. First, does case-specific AI assistance improve human discrimination between authentic and synthetic visual media? Second, how do explanation format, actual advice quality, and disclosed performance affect reliance? Third, what limits constrain transfer from these experiments to criminal evidence evaluation? The review develops an evidence-based account of these boundaries rather than proposing another unimplemented participant experiment.

Methods

Review design and reporting

This is a systematic review with a structured synthesis without meta-analysis. Reporting was organized with reference to PRISMA 2020, PRISMA-S, and SWiM (Page et al. 2021; Rethlefsen et al. 2021; Campbell et al. 2020). These are reporting frameworks, not certifications of methodological quality. The review was developed from an earlier narrative manuscript. Eligibility was operationalized during that development, and the review was not prospectively registered. The final search expressions, exported identifiers, duplicate links, screening decisions, and extraction fields are retained in the accompanying review materials.

Searches, preliminary screening, extraction, and initial critical appraisal were performed in a single AI-assisted workflow. No independent duplicate human screening or extraction was performed, and no inter-reviewer agreement statistic is claimed. AI assistance was also used for drafting and formatting. The absence of independent verification is a material review-level limitation. The reporting distinguishes the actual procedures from stronger procedures that would be desirable in a replicated review.

Eligibility criteria

Eligible reports described a primary comparative human study in which adults judged whether a facial image or visual video was authentic or AI-generated/manipulated, with case-specific AI assistance presented before the final judgment. Assistance could comprise a prediction, confidence score, saliency map, or detector-linked artifact amplification. Simulated AI advice was eligible because it can isolate the consequences of receiving an assessment, although it cannot validate a deployed detector. An objective classification, discrimination, or correctness-contingent reliance outcome was required.

Comparators could be an initial unassisted judgment, a separate unassisted condition, another assistance format, another advice source, or manipulated advice quality. Randomized, counterbalanced, and other comparative experimental arrangements were eligible. Journal papers, conference papers, and research preprints with an assessable full report were considered. English-language reports published or publicly available from 1 January 2018 through 20 October 2026 were eligible. Publication status was recorded rather than used as an automatic quality judgment.

Text-only and speech-only tasks were excluded to retain a defined visual-evidence question. General education, incentives, and warnings without case-specific AI assistance were outside the primary synthesis. Also excluded were technical benchmarks without assisted-human decisions, computational fusion of separately collected human and machine judgments, qualitative needs studies, and evaluations reporting only perceived clarity, trust, or usefulness. Secondary reviews were background sources rather than included primary studies. Relevant abstracts without an assessable full report were recorded as not retrieved and were not treated as completed evidence assessments.

Information sources and search strategy

Europe PMC and OpenAlex were queried through their public APIs on 20 October 2026. The Europe PMC expression combined title/abstract terms for deepfakes or synthetic faces, human participants, and assistance, advice, training, warnings, feedback, explanations, or interventions. The broad intervention block was intended to avoid missing reports whose abstracts used a neighboring term. The response returned 97 records, below the 1,000-record retrieval limit.

The OpenAlex search combined a title search for deepfake with a title/abstract search for human and at least one of guidance, advice, assistance, explanations, or collaboration, together with the same publication-date bounds. All 105 returned records were retrieved in one response, below the 200-record request limit. The exact API filters, their returned counts, and retrieval addresses are reproduced in Appendix A. Exploratory searches used while developing the topic were not represented as additional bibliographic-database exports.

Targeted citation follow-up from the retrieved human–AI literature supplied four further candidate reports: Boyd et al. (2023), Zhou et al. (2025), Pearson et al. (2026), and Fosco et al. (2025). Publisher and author-hosted pages were used to locate the reports and check bibliographic information. Scopus, Web of Science, PsycINFO, and subscription database interfaces were not searched. The resulting coverage is therefore bounded by two open indexes and the documented citation follow-up, not an exhaustive search of every scholarly database.

For the expanded introduction and discussion, targeted, non-systematic literature searches on 21 October 2026 identified technical benchmarks, automation research, forensic communication, legal scholarship, and review methodology. Eighty-one additional contextual sources were retained after checking bibliographic records against publisher or author-repository records; these sources were not counted as additional intervention studies. The comparative deepfake-assistance corpus, eligibility rules, and selection counts remain those of the systematic search described above. Contextual evidence is used to interpret mechanisms and propose safeguards, with cross-domain applications identified as inferences rather than synthesized intervention effects.

Selection and data collection

Records were deduplicated using normalized DOI and title matches. Normalization removed case and punctuation differences and DOI URL prefixes. Identical-title versions were consolidated, with the preferred report identified during retrieval. Distinct reports were not assumed independent merely because their titles differed. The screening unit was the bibliographic report, and the synthesis unit was the relevant assistance comparison within that report.

The 206 identified records yielded 182 unique records after 24 duplicates were removed. Title and available abstract screening excluded 170 records. Twelve reports were sought for full-text assessment; one could not be retrieved through the available routes. Eleven full reports were assessed, four were excluded, and seven were included. The unretrieved report and the four full-text exclusions are identified individually in Appendix B. Figure 1 presents these observed counts.

Extraction recorded publication status, media and participant setting, the eligible experimental component, comparison, real versus simulated assistance, communicated versus delivered performance, outcome definition, principal result, and limitations. The reported sample refers to the relevant component where available, not necessarily every experiment in the paper. No participant total was obtained by adding samples across reports. Related artifact-amplification reports were retained as reports but treated cautiously because shared materials and uncertain sample independence would make simple accumulation misleading.

Critical appraisal

A review-specific, domain-based appraisal examined five issues: comparability and allocation; isolation of assistance from practice or other changes; exclusions and missing observations; objective outcome measurement; and transparency of analyses and reporting. Each domain was recorded as a documented strength, a concern, or unclear, with a report-specific explanation. This was an interpretive appraisal, not a validated numerical scale or a formal RoB 2 assessment. No total quality score was calculated.

A separate applicability assessment considered the participant population, stimulus selection, generation method, advice realism, and distance from a contested legal decision. A study can offer a useful experimental contrast while having limited forensic applicability. Perfectly correct labels, for example, can test whether people respond to an interface without showing how that interface behaves when the underlying detection system is wrong. Applicability judgments were kept distinct from internal design concerns.

The appraisal distinguishes reporting completeness from risk of bias and legal applicability. RoB 2 and ROBINS-I illustrate design-specific approaches to intervention bias (Sterne et al. 2019; Sterne et al. 2016), but no completed assessment using either tool is claimed here. AMSTAR 2 and ROBIS address the quality or bias of reviews rather than the validity of a detector (Shea et al. 2017; Whiting et al. 2016). Citing these frameworks does not remedy the absence of duplicate human assessment; it clarifies which stronger procedures a future update would need.

Synthesis strategy

Reports were grouped by prediction and advice presentation, explanation or artifact amplification, and communicated versus delivered performance. Numerical findings were retained in their original units. Percentage-point changes were calculated only by subtracting clearly reported percentages within the same comparison. Confidence intervals, standardized effects, and significance levels were not reconstructed where the necessary information was unavailable. A result without a reported interval was not given an invented interval.

Meta-analysis was not conducted because the eligible contrasts differed in media, intervention timing, advice correctness, outcome definition, and dependence structure. Some comparisons involved repeated responses to the same item; others involved different interfaces or experimentally assigned labels. A common pooled effect would obscure which component had changed. The synthesis therefore evaluates the direction and conditions of a result without counting statistically significant findings as votes for effectiveness. No funnel plot, statistical publication-bias test, or formal GRADE assessment was undertaken.

Results

Study selection and characteristics

Seven reports met the eligibility criteria: Groh et al. (2022), Boyd et al. (2023), Josephs et al. (2024), Fosco et al. (2025), Zhou et al. (2025), Pearson et al. (2026), and Kramer and Lister (2026). Six were journal or conference publications; the retrieved Fosco report was an updated research preprint. Three concerned static facial images and four concerned video. Several papers contained additional experiments not used to answer the assistance question.

Table 1 identifies the eligible component and principal interpretive limit for each report. These studies were primarily general-population or convenience online tasks, not trials of judges deciding contested admissibility or guilt. No eligible study established effects on criminal verdict accuracy, attribution to a defendant, or access to effective challenge. Those outcomes remain an evidence gap within this review’s search and eligibility boundaries.

Counts refer to retrieved records and reports, not estimated search-engine totals. The flow follows the review ledger.

Table 1
Included reports and eligible assistance comparisons

Report	Media and sample	Assistance comparison	Principal finding	Main limit
Groh et al. (2022)	Video; 304 recruited in the relevant cohort	Initial judgment followed by a real model prediction	Accuracy 66% → 73%; incorrect advice could harm	Sequential decisions; selected clips
Boyd et al. (2023)	Faces; about 1,560 across study components	Labels, saliency maps; actual/reported reliability	Labels generally more useful than saliency maps	Phase and component samples differ
Josephs et al. (2024)	Video; 298 in Study 2	Text versus artifact-amplification display	Greater uptake and confidence with amplification	Assessment labels were correct
Fosco et al. (2025)	Video; 30 per group in the stated comparison	Unaided, text flag, caricature flag	Deepfake hit rates 0.53, 0.78, 0.94	Preprint; related research materials
Zhou et al. (2025)	Faces; 390 analyzed	Stated FPR 3% versus 30%; simulated advice	Reliance depended on disclosure and prediction	Stated FPR did not match displayed errors
Pearson et al. (2026)	Faces; 295 analyzed	Matched guidance attributed to AI or humans	No overall accuracy difference by advice source	Advice correct on half the trials
Kramer and Lister (2026)	Video; 132 and 139 in eligible experiments	Simulated guidance; quality and prevalence varied	Higher-quality guidance helped; bad advice misled	Convenience sample; notable exclusions

Prediction assistance and advice-source effects

In the recruited video cohort of Groh et al. (2022), showing model predictions increased correct identification from 66% to 73%. Incorrect or equivocal predictions nevertheless reduced performance on affected items. The contrast demonstrates a potential average benefit together with item-level vulnerability; it does not show that every assisted decision improved.

Boyd et al. (2023) compared labels, saliency maps, and combinations in synthetic-face judgments, and also varied actual and reported AI accuracy. Decision-only support was more useful than saliency maps in their setting. For a 90%-accurate advisor, reported participant accuracy was 62.36% when the advisor was described as 60% accurate and 65.15% when described as 90% accurate. This descriptive contrast shows that disclosure and delivered performance can be manipulated separately; the advisor's accuracy is not the final human accuracy.

Pearson et al. (2026) compared identical guidance attributed to AI or humans. Advice was correct on half the trials. Mean classification accuracy was approximately 66% with AI guidance and 67% with

human guidance, without a significant source difference. More favorable AI attitudes were associated with poorer discrimination in the AI condition. The result qualifies a general claim that the AI label alone reliably improves or harms performance.

Explanation and artifact-amplification formats

Josephs et al. (2024) compared text prompts with artifact amplification in a 298-participant assistance experiment. The latter increased uptake of the presented assessment and confidence on more difficult material. Crucially, the labels followed ground truth. The study therefore informs communication of correct assessments but does not establish resilience to incorrect detector advice.

The retrieved Fosco et al. (2025) preprint compared unaided viewing, textual flags, and caricature-based flags, with 30 participants per condition in that comparison. Reported deepfake hit rates were 0.53, 0.78, and 0.94, respectively. These are detection rates for synthetic items, not an overall judicial accuracy estimate. The related artifact-amplification reports were not combined into an independent pooled estimate.

The eligible formats demonstrate that assistance need not be a verbal claim. Altering the visible stimulus can make an assessment compelling by changing what the observer sees. This creates a distinct evidentiary question: whether a transformed explanatory exhibit makes the original file easier to evaluate, or makes the transformation itself appear to prove manipulation. That distinction is an implication drawn here, not an outcome tested in the included experiments.

Error information and delivered advice quality

Zhou et al. (2025) analyzed 390 participants and described notional advisors as having false-positive rates of 3% or 30%. The displayed predictions did not instantiate those rates. Trust was slightly higher for the lower stated rate; the reported standardized difference was small ($d = 0.09$). Reliance varied with the prediction and risk framing. Consequently, the experiment tests responses to communicated performance under simulated advice rather than a real reduction in detector error.

Kramer and Lister (2026) examined simulated video guidance in Experiments 3 and 4, with analyzed samples of 132 and 139. In Experiment 3, initial-to-final accuracy changes were approximately one, 12, and 31 percentage points under 60%, 80%, and 100% accurate guidance. The one-point change was not statistically significant. Participants also moved toward incorrect advice. The experiment separates benefit from advice quality from any claim that human oversight consistently corrects algorithmic errors.

These comparisons do not supply a universal optimal disclosure format. Global accuracy, a conditional false-positive rate, and a case-specific prediction communicate different quantities. The available experiments also differ in whether displayed reliability matches the advice participants experience. Their results should therefore remain attached to the manipulated information rather than be summarized as a single effect of “transparency.”

Appraisal and consistency of the evidence

Table 2 records the principal appraisal concerns. Ground-truth classification was an important strength across the included reports. However, before–after arrangements can combine advice exposure with repetition, and reporting did not always allow a clean assessment of allocation concealment or all analytic choices. These concerns limit how confidently an observed difference can be assigned to one interface component.

Table 2
Structured appraisal of the eligible comparisons

Report	Documented strength	Concern for interpretation	Applicability limit
Groh et al.	Objective truth and item-level error analysis	Initial/final sequence; nonrecruited self-selection	Benchmark video task
Boyd et al.	Actual and reported reliability distinguished	Allocation detail not uniformly clear across components	Synthetic faces and nonexpert users
Josephs et al.	Preregistered exclusions and objective measures	Correct-label design cannot estimate bad-advice harms	Interface use under selected conditions
Fosco et al.	Controlled comparison and stated group sizes	Preprint; report linkage and selective flagging	No independent forensic validation of display
Zhou et al.	Randomized order and stated exclusion count	Risk manipulation unsuccessful; notional error rates	Communication response, not detector validation
Pearson et al.	Preregistration and counterbalanced advice streams	Unassisted reference is external; selected difficulty	General-population image decisions
Kramer and Lister	Controlled advice correctness; sensitivity checks	Large exclusion counts and unclear allocation sequence	Mobile task and convenience recruitment

The evidence supports a conditional account of assistance rather than a uniform direction of effect. An advisor may correct human errors, reproduce them, or introduce new ones. Information that encourages uptake can improve the first situation and worsen the third. A numerical summary that averages over advice correctness can conceal this distinction. Likewise, improvement against an unassisted baseline is different from improvement against the advisor operating alone. The review did not find a basis for treating these comparisons as interchangeable.

Discussion

What the systematic comparison adds

The main contribution is the separation of three questions commonly collapsed in discussions of forensic AI: whether the detector classifies media accurately, whether a person changes a decision after seeing its output, and whether the resulting decision is better. Each requires a different comparison. Detector benchmarking answers the first; assisted-human experiments address the second and third. A preferred explanation or higher trust score can be relevant to usability while remaining insufficient evidence of improved evidentiary judgment.

This distinction also explains why a systematic review can identify a modest set of relevant experiments within a much larger technical literature. The exclusion of independent machine benchmarks was substantive rather than dismissive. Such benchmarks remain essential for validating a tool. They were excluded because they cannot show how a recipient interprets and uses its opinion. Conversely, the included simulated-advice experiments can identify communication effects without demonstrating that any available system achieves the stated reliability.

The synthesis suggests evaluating assistance at two levels. At the system level, the question is whether final judgments improve across a defined task distribution. At the case level, the question is whether reliance discriminates between helpful and misleading advice. A tool can benefit the average user while harming a subset of difficult cases. In criminal evidence, the individual disputed file may belong to that subset. Average performance must therefore be accompanied by an account of where the assistance fails and whether recipients can recognize those failures.

Technical validity and transport to disputed evidence

A detector's performance depends on what it has learned to recognize. MesoNet examines mesoscopic image properties, while face-warping detection targets artifacts associated with a particular generation and transformation process (Afchar et al. 2018; Li and Lyu 2018). Face X-Ray instead emphasizes blending boundaries (Li et al. 2020). These are distinct evidentiary hypotheses. A positive result should identify the relevant trace and tested manipulation family; it should not be translated automatically into a universal claim that a recording is fabricated.

Generalization also depends on the relationship between training and deployment. Studies of CNN-generated images and frequency-domain artifacts show that shared generation features can support detection across some tested sources (Wang et al. 2020; Frank et al. 2020). However, shortcut learning explains why a rule can perform well under benchmark conditions yet fail when incidental cues change (Geirhos et al. 2020). The practical inference is to test the submitted media's processing conditions, including compression and resizing, rather than assume that a published aggregate score transfers unchanged to casework.

Documentation can make these limits inspectable. Model cards and dataset datasheets describe intended uses, evaluation conditions, and data provenance (Mitchell et al. 2019; Gebru et al. 2021). A lifecycle account of machine-learning harm further directs attention to data collection, development, and deployment rather than locating every problem in the final classifier (Suresh and Guttag 2021). Applied

here, these frameworks support a version-specific validation record, explicit unsupported uses, and a record of preprocessing. They provide documentation principles, not evidence that a completed reporting form prevents wrongful judgments.

Reliance as a conditional decision process

The relevant design question is which human function receives support. Automation can change information acquisition, analysis, decision selection, or implementation (Parasuraman et al. 2000). Its misuse and disuse are separate possibilities (Parasuraman and Riley 1997). A deepfake display that highlights a suspicious region changes inspection; a binary authenticity label proposes a conclusion. Their effects should be tested separately because they distribute attention, responsibility, and opportunities for independent checking differently.

Automation-bias research distinguishes failures to act when a system omits a cue from acceptance of an erroneous recommendation (Skitka et al. 1999; Mosier et al. 1998). Trust research also cautions that reliance is not adequately captured by a single favorable attitude measure (Dzindolet et al. 2003). In this review's setting, agreement with incorrect AI is therefore especially informative: it can reveal whether assistance replaces inspection. A system that raises overall accuracy may still increase precisely this form of error on the subset of cases where its own judgment fails.

Trust has dispositional, situational, and learned components, while research on human–robot interaction links trust to characteristics of the system and its performance (Hoff and Bashir 2015; Hancock et al. 2011). These literatures help explain why a numerical error disclosure may interact with experience and task demands. They do not establish the magnitude of any courtroom effect. A legal user repeatedly exposed to easy, correct recommendations may develop expectations that are poorly calibrated for a later difficult or unfamiliar file.

Neither blanket skepticism nor automatic acceptance is the appropriate objective. Algorithm-aversion research documents avoidance after observed mistakes, whereas algorithm-appreciation research identifies conditions favoring algorithmic advice (Dietvorst et al. 2015; Logg et al. 2019). Perceived task objectivity can also influence willingness to rely (Castelo et al. 2019). Accordingly, legal framing may affect uptake independently of technical validity. The experiment should ask whether participants distinguish sound from unsound advice, not merely whether an interface increases acceptance of a technology labeled objective.

Verification imposes its own burden. Reviews of automation bias identify task complexity, cognitive demands, and system design as relevant conditions (Goddard et al. 2012; Lyell and Coiera 2017). In forensic work, a user may possess the authority to reject an output yet lack the information or time needed to test it. Meaningful oversight therefore requires an attainable checking task. Asking an examiner to remain responsible while withholding validation data risks making oversight a formal allocation of blame rather than a demonstrated source of accuracy.

Interventions should be evaluated against this checking burden. Cognitive forcing can reduce overreliance in other decision tasks, but can also impose usability costs (Buçinca et al. 2021). Explanations have not consistently increased complementary team performance, and confidence-based trust calibration does not guarantee improved joint decisions (Bansal et al. 2021; Zhang et al. 2020). A distinction between warranted trust and merely increased trust is therefore necessary (Jacovi et al. 2021). The present synthesis supports testing these mechanisms with deepfake evidence; it does not establish that slowing users down or adding explanations will reliably protect criminal defendants.

From transparency to contestable assessment

Transparency should expose an assessment to scrutiny. It should identify the tested proposition, input requirements, validation population, decision threshold, model version, and relevant errors. A readable explanation is helpful only to the extent that it conveys information supported by the method. An intuitively persuasive visual region or fluent narrative can otherwise become an additional assertion requiring validation.

The practical reporting unit should be a bounded claim. “The file contains features associated with a specified manipulation under the tested conditions” is different from “the recording is false” and much further from “the defendant fabricated it.” A report should identify which proposition the method can support and which require other evidence. This recommendation follows from the structure of inference; the included tasks did not test whether adding such a statement improves courtroom reasoning.

Contestability also requires preserving the relationship between an original file and any explanatory transformation. An artifact-amplified video should be labeled as a derived display, with the original retained and the transformation documented. A recipient must be able to ask whether the same processing would produce suspicious features in authentic media. The evidentiary value cannot rest solely on the observation that a tool designed to exaggerate irregularities displays an irregularity.

Explanation methods also answer different questions. Local surrogate explanations and additive feature attributions describe aspects of a model's prediction behavior (Ribeiro et al. 2016; Lundberg and Lee 2017). They do not, without additional validation, establish that the highlighted property is a manipulation artifact or that a particular person created it. For an evidentiary report, the explanation's target must be explicit: model behavior, a property of the file, the generation mechanism, and the actor's responsibility are separate propositions.

Interpretability therefore needs an evaluation standard. A rigorous science of interpretation requires specifying what an explanation is intended to achieve, while critiques of interpretability warn against treating the term as a single measurable property (Doshi-Velez and Kim 2017; Lipton 2018). The argument for interpretable models in high-stakes decisions provides a further alternative to assuming that a post-hoc explanation repairs an opaque model (Rudin 2019). These positions justify testing understandable alternatives where feasible; they do not establish that a simpler detector is necessarily accurate for the media at issue.

Visual plausibility is an insufficient fidelity test. Saliency-map sanity checks show why an attractive explanation can fail to depend appropriately on a trained model or its data (Adebayo et al. 2018). Related criticism in health-care AI emphasizes rigorous validation rather than explanation as a substitute for validity (Ghassemi et al. 2021). Transferring that caution to deepfake evidence suggests separately testing whether a display faithfully represents the method and whether recipients understand its limits. A polished demonstrative exhibit should not acquire additional probative weight merely because its design is persuasive.

Conditional errors and case-specific conclusions

A false-positive rate is the proportion of authentic material classified as synthetic within a defined validation setting. It is not the probability that a positive result is wrong. A model score attached to a file is also not automatically a calibrated posterior probability. The relevant reference population, prevalence, and relationship between the submitted file and the validation material affect the permissible interpretation.

An arithmetic illustration makes the distinction concrete. A hypothetical binary detector with 90% sensitivity and 95% specificity has a positive predictive value of approximately 67% when synthesis prevalence is 10%, but approximately 15% when prevalence is 1%. These values are calculations under stated assumptions, not extracted experimental outcomes or performance claims about a real detector. They demonstrate why conditional denominators and population assumptions must be visible.

Disclosure

may therefore require more than reporting a small error number. Readers need to know which class forms the denominator, how the reference population was assembled, and whether the estimate applies to the media at issue. Where prevalence cannot be justified, a report should not manufacture a case-specific probability. Presenting sensitivity to several transparent assumptions may be more informative than a single reassuring percentage.

Adjacent forensic-communication research reinforces the need to preserve domain boundaries. Garrett et al. (2020) examined error-rate and likelihood-ratio qualifications for fingerprint and voice evidence, finding different responses across the two disciplines. This is contextual literature rather than an included deepfake-assistance experiment. It supports asking whether a disclosure works for the particular forensic task, while supplying no direct estimate of its effect on deepfake judgments. A reform based on cross-domain analogy should make that evidentiary gap explicit.

The danger of reversing conditional probabilities has a long forensic history. The prosecutor's fallacy concerns confusion between a probability conditional on a hypothesis and the probability of that hypothesis given the evidence (Thompson and Schumann 1987). Base-rate reasoning also depends on the reliability and structure of the reference information, rather than on mechanically inserting a convenient prevalence (Koehler 1996). A detector's false-positive rate consequently cannot establish the

chance that a defendant fabricated a recording, nor can it supply missing evidence connecting the defendant to its production.

Natural frequencies offer a testable communication alternative. Research on frequency formats and a subsequent meta-analysis support their usefulness for many Bayesian reasoning tasks, while identifying important moderators (Gigerenzer and Hoffrage 1995; McDowell and Jacobs 2017). A report might describe expected true and false alerts among an explicitly hypothetical number of files under stated assumptions. The exercise should expose the denominator and uncertainty, not imply that the hypothetical collection represents the actual criminal-case population. Whether this format improves evaluation of deepfake evidence remains an empirical question.

Verbal labels also require testing. Research on uncertain forensic opinion and lay understanding of forensic statistics shows that evidence strength, format, prior expectations, and discipline can affect interpretation (Martire et al. 2013; Thompson and Newman 2015). Experiments on communicating numerical uncertainty outside criminal justice provide additional reason not to assume that acknowledging uncertainty necessarily destroys trust (van der Bles et al. 2020). The appropriate question is whether communication improves understanding and discrimination. A modest reduction in confidence can be desirable if it corrects an exaggerated belief in what the evidence proves.

Relevance to criminal evidence evaluation

The included experiments concern authenticity judgments. Their legal relevance is indirect: they identify ways in which recipients may respond to an assessment that could later enter an investigative or adjudicative process. They do not establish a rule of admissibility, an acceptable legal error threshold, or a probability of guilt. Any institutional recommendation must retain that boundary.

A criminal-evidence workflow involves additional questions about acquisition, provenance, alteration history, chain of custody, and alternative explanations. Technical classification addresses only part of that process. An authentic recording may support a misleading narrative; synthetic material may have a legitimate explanatory purpose. Conversely, a failed detection may reflect the tool's limits rather than proof of authenticity. These possibilities make corroboration and a clearly stated inferential scope necessary even when an authenticity assessment is technically competent.

The opportunity to challenge an assessment should encompass the information needed to understand its limits. Depending on the setting, that may involve the original file, preprocessing history, model identifier, validation report, threshold, and examiner interventions. This review does not establish that disclosure of every item improves lay comprehension. Its narrower implication is that confidence or willingness to rely cannot substitute for access to the basis of the conclusion.

Cognitive independence and effective challenge

A human examiner is not an independent safeguard merely by remaining in the workflow. Studies of fingerprint identification and DNA mixture interpretation illustrate susceptibility to contextual information

in different forensic domains (Dror et al. 2006; Dror and Hampikian 2011). The broader forensic-confirmation-bias literature and work on expert cognition show why technical experience does not by itself eliminate this risk (Kassin et al. 2013; Dror 2020). These findings do not quantify bias in deepfake examination. They identify a plausible pathway through which an AI label can become an anchor for subsequent visual interpretation.

The order of information exposure is consequently a substantive variable. Forensic examiners' reported beliefs about their own immunity to bias provide little basis for relying on self-assurance alone (Kukucka et al. 2017). Target shifting after a suspect's profile becomes known illustrates another route to distorted interpretation (Thompson 2009). Context-management proposals, including sequential unmasking, and earlier analysis of observer effects suggest recording an initial assessment before exposing the examiner to potentially biasing information (Dror et al. 2015; Risinger et al. 2002). This is a proposed design safeguard whose benefits and costs require direct testing in deepfake casework.

Independent challenge concerns access as well as explanation. Scholarship on accountable algorithms and counterfactual explanations offers different ways to make automated decisions inspectable (Kroll et al. 2017; Wachter et al. 2018). In criminal proceedings, however, a counterfactual account of which input would change a label may leave the validity of the label unresolved. Wexler's analysis of criminal trade secrets highlights the institutional obstacle when developers resist disclosure (Wexler 2018). A clear interface cannot compensate for inability to inspect the actual method, inputs, and relevant validation underlying an accusation.

Machine-evidence scholarship further distinguishes technical output from the safeguards needed to assess its credibility (Roth 2017; Roth 2016). Analysis of AI as evidence likewise foregrounds validity, reliability, testing, and the danger of extending a tool beyond its supported purpose (Grimm et al. 2021). These are legal and normative frameworks, not intervention estimates from this review. Their application is that a proponent should identify the precise claim the system supports and permit scrutiny of the implementation used in the case, including material departures from validated conditions.

The forensic-science reform literature supplies a further institutional lesson: experience and longstanding acceptance cannot replace empirical validation (Saks and Koehler 2005; Mnookin et al. 2011; Saks and Faigman 2008). Research on invalid forensic testimony in DNA-exoneration cases illustrates the consequences of overstating scientific support, without providing a representative error rate for all forensic testimony (Garrett and Neufeld 2009). Forensic AI should therefore be evaluated against a documented claim rather than a presumption that either automation or expert endorsement guarantees scientific reliability.

Validation must also reflect the conditions of use. Work on assessing forensic accuracy argues for tests capable of measuring performance under appropriate conditions (Koehler 2017). Digital-evidence practice emphasizes the recovery and interpretation of traces within an investigation, while a broader forensic-science perspective situates those traces within the crime and its reconstruction (Casey 2011; Roux et al. 2012). For deepfake evidence, these approaches support preserving provenance and

alternative explanations alongside detector output. Authenticity classification remains one component of a reconstruction, not a complete account of an event or its author.

Research priorities for experimental criminology

Future studies should distinguish actual error performance from descriptions of performance. A factorial design can vary the quality of the advice, its stated reliability, and its presentation while maintaining a constant underlying evidence set. Correct and incorrect recommendations should occur on both authentic and synthetic material. Otherwise, an interface optimized to make correct advice persuasive remains untested against its most consequential failure condition.

Outcomes should include sensitivity, specificity, response criterion, uncertainty calibration, and acceptance of incorrect advice. Confidence, perceived usefulness, and overall agreement can be secondary measures. Investigators should also measure inference beyond authenticity: for example, whether a participant attributes production or criminal intent to a person without supporting evidence. Such outcomes would connect laboratory authenticity tasks more directly with substantive criminal-justice questions.

Professional populations should be studied without treating legal experience as equivalent to technical expertise. Judges, advocates, investigators, and forensic practitioners may have different knowledge and different responsibilities. Designs should preserve these distinctions, document the decision being made, and consider cross-examination or competing expert reports. Simulations with convenience samples remain useful, but claims about improved legal practice require evidence from those additional conditions.

A further priority is independent validation of explanatory displays. A transformed image or narrative should be tested for both its correspondence to the method and its effect on recipients. Explanatory fidelity and communication effectiveness are separate outcomes. A persuasive display that accurately communicates a weak or inapplicable method still cannot establish the underlying conclusion. Robust evaluations should include cases where the method is confident but wrong.

Future experiments should sample media as deliberately as participants. Work on stimulus sampling shows why increasing participant numbers alone may not resolve dependence on an idiosyncratic stimulus set (Westfall et al. 2015). The generalizability critique extends this concern to tasks and settings that a statistical model does not adequately represent (Yarkoni 2022). Replications should therefore vary generators, recording conditions, languages, and case contexts while reporting which generalizations their sampling and analysis actually support.

Claims of no meaningful difference require more than a nonsignificant test. Equivalence testing offers one framework when a justified smallest effect of interest has been specified (Lakens 2017). Conversely, quantitative synthesis needs a defensible common estimand; measuring inconsistency does not make substantively different contrasts interchangeable (Higgins et al. 2003). Selective-reporting diagnostics

such as p-curve have their own evidential requirements and were not applied here (Simonsohn et al. 2014). Preregistered comparisons, full outcome reporting, and accessible materials would provide a firmer basis for future cumulative assessment.

Limitations

This review has several limitations. The indexed searches used Europe PMC and OpenAlex, which have different coverage and metadata completeness. The OpenAlex title requirement could miss studies using terms such as synthetic faces without deepfake in the title; targeted citation follow-up recovered some such reports but cannot guarantee complete recovery. Non-English reports and subscription database searches were outside the performed search. The seven reports should therefore be understood as the included evidence located by this documented process.

Screening and extraction were performed in one AI-assisted stream without independent duplicate human review. Classification errors, missed studies, and interpretive errors remain possible. The review-specific appraisal is less standardized than a validated risk-of-bias procedure and should not be mistaken for one. The search strategy was developed while converting an existing narrative manuscript rather than prospectively registered before all topic exploration.

The synthesis was limited by heterogeneous outcomes and experimental arrangements. It did not estimate a pooled effect, formally assess publication bias, or grade certainty using GRADE. A small set of reports also cannot establish how effects vary across every population, generator, and platform. Different reports may share datasets or research lineages, so their findings were not counted as statistically independent replications.

One potentially relevant report was not available for full assessment. Its abstract was not used to fill missing methods or extract unverified estimates. More generally, absence of an eligible study in this synthesis is not proof that no such study exists. In particular, the conclusion about the lack of direct verdict or criminal-attribution outcomes applies to the reports included here.

Conclusions

Case-specific AI assistance can change how people judge visual deepfakes, but its value depends on advice quality, presentation, and the outcome assessed. Improved average classification, greater confidence, and increased agreement with AI are distinct results. The reviewed evidence supports testing them separately and examining what happens when the assessment is incorrect.

For criminal evidence, the central implication is a requirement for bounded and contestable conclusions. Error information should identify its denominator and validation conditions, explanatory transformations should be documented, and authenticity should remain separate from attribution and guilt. The included experiments inform these concerns while leaving direct effects on legal decision-making unresolved. Better forensic communication remains a testable safeguard rather than an established substitute for technical validation and procedural scrutiny.

Declarations

Funding No external funding was received for preparation of this article.

Competing interests The author declares no competing interests.

Ethics approval and consent Not applicable. The review used published reports and public bibliographic metadata and involved no new participant recruitment.

Data and materials availability The accompanying review materials contain the identification and screening ledgers, search expressions, extraction matrix, appraisal notes, and search-retrieval code. No individual participant data were obtained or reanalyzed. No public repository registration is claimed.

Use of generative AI ChatGPT (OpenAI) was used for searches, single-stream preliminary screening, extraction, initial appraisal, drafting, language editing, and formatting. Independent duplicate human verification was not undertaken. The author retains responsibility for reviewing the sources, decisions, and final content. AI is not an author.

References

1. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B (2018) Sanity checks for saliency maps. *Advances in Neural Information Processing Systems, 31*. https://papers.neurips.cc/paper_files/paper/2018/hash/294a8ed24b1ad22ec2e7efea049b8737-Abstract.html
2. Afchar D, Nozick V, Yamagishi J, Echizen I (2018) MesoNet: A compact facial video forgery detection network. *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, 1–7. https://doi.org/10.1109/wifs.2018.8630761
3. Bansal G, Wu T, Zhou J, Fok R, Nushi B, Kamar E, Ribeiro MT, Weld D (2021) Does the whole exceed its parts? The effect of AI explanations on complementary team performance. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3411764.3445717
4. Boyd A, Tinsley P, Bowyer K, Czajka A (2023) The Value of AI Guidance in Human Examination of Synthetically-Generated Faces. *Proceedings of the AAAI Conference on Artificial Intelligence*, *37*(5), 5930–5938. https://doi.org/10.1609/aaai.v37i5.25734
5. Bray SD, Johnson SD, Kleinberg B (2023) Testing human ability to detect 'deepfake' images of human faces. J Cybersecur 9(1):tyad011. https://doi.org/10.1093/cybsec/tyad011
6. Buçinca Z, Malaya MB, Gajos KZ (2021) To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW1), 1–21. https://doi.org/10.1145/3449287
7. Campbell M, McKenzie JE, Sowden A, Katikireddi SV, Brennan SE, Ellis S, Hartmann-Boyce J, Ryan R, Shepperd S, Thomas J, Welch V, Thomson H (2020) Synthesis without meta-analysis (SWiM) in

systematic reviews: reporting guideline. BMJ 368:l6890. https://doi.org/10.1136/bmj.l6890

8. Casey E (2011) *Digital evidence and computer crime: Forensic science, computers, and the internet (3rd ed.). Academic Press*. https://shop.elsevier.com/books/digital-evidence-and-computer-crime/casey/978-0-08-092148-8
9. Castelo N, Bos MW, Lehmann DR (2019) Task-dependent algorithm aversion. J Mark Res 56(5):809–825. https://doi.org/10.1177/0022243719851788
10. Castro F, Gattulli V, Impedovo D, Monaco A (2026) DeepDect: an explainable AI platform for face swapping and face generation DeepFake detection. Multimedia Tools Appl 85(5):500. https://doi.org/10.1007/s11042-026-21672-1
11. Chesney R, Citron DK (2019) Deep fakes: A looming challenge for privacy, democracy, and national security. Calif Law Rev 107(6):1753–1820. https://law.utexas.edu/faculty/publications/2019-deep-fakes-a-looming-challenge-for-privacy-democracy-and-national-security/
12. Cooke D, Edwards A, Barkoff S, Kelly K (2025) As good as a coin toss: Human detection of AI-generated content. Commun ACM 68(10):100–109. https://doi.org/10.1145/3729417
13. Diel A, Lalgi T, Schröter IC, MacDorman KF, Teufel M, Bäuerle A (2024) Human performance in detecting deepfakes: A systematic review and meta-analysis of 56 papers. Computers Hum Behav Rep 16:100538. https://doi.org/10.1016/j.chbr.2024.100538
14. Dietvorst BJ, Simmons JP, Massey C (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. J Exp Psychol Gen 144(1):114–126. https://doi.org/10.1037/xge0000033
15. Dolhansky B, Bitton J, Pflaum B, Lu J, Howes R, Wang M, Canton Ferrer C (2020) The DeepFake detection challenge (DFDC) dataset [preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2006.07397
16. Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning [preprint]. *arXiv*. https://doi.org/10.48550/arXiv.1702.08608
17. Dror IE (2020) Cognitive and human factors in expert decision making: Six fallacies and the eight sources of bias. Anal Chem 92(12):7998–8004. https://doi.org/10.1021/acs.analchem.0c00704
18. Dror IE, Hampikian G (2011) Subjectivity and bias in forensic DNA mixture interpretation. Sci Justice 51(4):204–208. https://doi.org/10.1016/j.scijus.2011.08.004
19. Dror IE, Charlton D, Péron AE (2006) Contextual information renders experts vulnerable to making erroneous identifications. Forensic Sci Int 156(1):74–78. https://doi.org/10.1016/j.forsciint.2005.10.017
20. Dror IE, Thompson WC, Meissner CA, Kornfield I, Krane D, Saks M, Risinger M (2015) Letter to the editor— context management toolbox: A linear sequential unmasking (LSU) approach for minimizing cognitive bias in forensic decision making. J Forensic Sci 60(4):1111–1112. https://doi.org/10.1111/1556-4029.12805
21. Dzindolet MT, Peterson SA, Pomranky RA, Pierce LG, Beck HP (2003) The role of trust in automation reliance. Int J Hum Comput Stud 58(6):697–718. https://doi.org/10.1016/s1071-5819(03)00038-7

22. Fosco C, Josephs E, Andonian A, Oliva A (2025) Deepfake caricatures: Amplifying attention to artifacts increases deepfake detection by humans and machines (Version 4) [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2206.00535
23. Frank J, Eisenhofer T, Schönherr L, Fischer A, Kolossa D, Holz T (2020) Leveraging frequency analysis for deep fake image recognition. *Proceedings of Machine Learning Research, 119*, 3247–3258. https://proceedings.mlr.press/v119/frank20a.html
24. Garrett BL, Neufeld PJ (2009) Invalid forensic science testimony and wrongful convictions. Va Law Rev 95(1):1–97. https://virginialawreview.org/articles/invalid-forensic-science-testimony-and-wrongful-convictions/
25. Garrett BL, Crozier WE, Grady R (2020) Error Rates, Likelihood Ratios, and Jury Evaluation of Forensic Evidence. J Forensic Sci 65(4):1199–1209. https://doi.org/10.1111/1556-4029.14323
26. Gebru T, Morgenstern J, Vecchione B, Vaughan JW, Wallach H, Daumé H III, Crawford K (2021) Datasheets for datasets. Commun ACM 64(12):86–92. https://doi.org/10.1145/3458723
27. Geirhos R, Jacobsen J-H, Michaelis C, Zemel R, Brendel W, Bethge M, Wichmann FA (2020) Shortcut learning in deep neural networks. Nat Mach Intell 2(11):665–673. https://doi.org/10.1038/s42256-020-00257-z
28. Ghassemi M, Oakden-Rayner L, Beam AL (2021) The false hope of current approaches to explainable artificial intelligence in health care. Lancet Digit Health 3(11):e745–e750. https://doi.org/10.1016/s2589-7500(21)00208-9
29. Gigerenzer G, Hoffrage U (1995) How to improve Bayesian reasoning without instruction: Frequency formats. Psychol Rev 102(4):684–704. https://doi.org/10.1037/0033-295x.102.4.684
30. Goddard K, Roudsari A, Wyatt JC (2012) Automation bias: A systematic review of frequency, effect mediators, and mitigators. J Am Med Inform Assoc 19(1):121–127. https://doi.org/10.1136/amiajnl-2011-000089
31. Grimm PW, Grossman MR, Cormack GV (2021) Artificial intelligence as evidence. Northwest J Technol Intellect Property 19(1):9–106. https://scholarlycommons.law.northwestern.edu/njtip/vol19/iss1/2/
32. Groh M, Epstein Z, Firestone C, Picard R (2022) Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*, *119*(1), e2110013119. https://doi.org/10.1073/pnas.2110013119
33. Groh M, Sankaranarayanan A, Singh N, Kim DY, Lippman A, Picard R (2024) Human detection of political speech deepfakes across transcripts, audio, and video. Nat Commun 15(1):7629. https://doi.org/10.1038/s41467-024-51998-z
34. Hancock PA, Billings DR, Schaefer KE, Chen JYC, de Visser EJ, Parasuraman R (2011) A meta-analysis of factors affecting trust in human-robot interaction. Hum Factors: J Hum Factors Ergon Soc 53(5):517–527. https://doi.org/10.1177/0018720811417254
35. Higgins JPT, Thompson SG, Deeks JJ, Altman DG (2003) Measuring inconsistency in meta-analyses. BMJ 327(7414):557–560. https://doi.org/10.1136/bmj.327.7414.557

36. Hoff KA, Bashir M (2015) Trust in automation: Integrating empirical evidence on factors that influence trust. Hum Factors: J Hum Factors Ergon Soc 57(3):407–434. https://doi.org/10.1177/0018720814547570
37. Jacovi A, Marasović A, Miller T, Goldberg Y (2021) Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 624–635. https://doi.org/10.1145/3442188.3445923
38. Josephs E, Fosco C, Oliva A (2024) Effects of Browsing Conditions and Visual Alert Design on Human Susceptibility to Deepfakes. J Online Trust Saf 2(2). https://doi.org/10.54501/jots.v2i2.144
39. Kassin SM, Dror IE, Kukucka J (2013) The forensic confirmation bias: Problems, perspectives, and proposed solutions. J Appl Res Memory Cognition 2(1):42–52. https://doi.org/10.1016/j.jarmac.2013.01.001
40. Köbis NC, Doležalová B, Soraperra I (2021) Fooled twice: People cannot detect deepfakes but think they can. iScience 24(11):103364. https://doi.org/10.1016/j.isci.2021.103364
41. Koehler JJ (1996) The base rate fallacy reconsidered: Descriptive, normative, and methodological challenges. Behav Brain Sci 19(1):1–17. https://doi.org/10.1017/s0140525x00041157
42. Koehler JJ (2017) Forensics or fauxrensics? Ascertaining accuracy in the forensic sciences. Arizona State Law J 49:1369–1416. https://arizonastatelawjournal.org/2018/02/07/forensics-or-fauxrensics-ascertaining-accuracy-in-the-forensic-sciences/
43. Kramer RSS, Lister VPM (2026) Improving human performance in deepfake detection: limits of aggregation and human–AI collaboration. Cogn Research: Principles Implications. https://doi.org/10.1186/s41235-026-00758-2
44. Kroll JA, Huey J, Barocas S, Felten EW, Reidenberg JR, Robinson DG, Yu H (2017) Accountable algorithms. Univ Pa Law Rev 165(3):633–705. https://pennlawreview.com/2017/02/23/accountable-algorithms/
45. Kukucka J, Kassin SM, Zapf PA, Dror IE (2017) Cognitive bias and blindness: A global survey of forensic science examiners. J Appl Res Memory Cognition 6(4):452–459. https://doi.org/10.1016/j.jarmac.2017.09.001
46. Lakens D (2017) Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. Social Psychol Personality Sci 8(4):355–362. https://doi.org/10.1177/1948550617697177
47. Lee JD, See KA (2004) Trust in Automation: Designing for Appropriate Reliance. Hum Factors: J Hum Factors Ergon Soc 46(1):50–80. https://doi.org/10.1518/hfes.46.1.50_30392
48. Li L, Bao J, Zhang T, Yang H, Chen D, Wen F, Guo B (2020) Face X-Ray for more general face forgery detection. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5000–5009. https://doi.org/10.1109/cvpr42600.2020.00505
49. Li Y, Lyu S (2018) Exposing DeepFake videos by detecting face warping artifacts [preprint]. *arXiv*. https://doi.org/10.48550/arXiv.1811.00656
50. Lipton ZC (2018) The mythos of model interpretability. Commun ACM 61(10):36–43. https://doi.org/10.1145/3233231

51. Logg JM, Minson JA, Moore DA (2019) Algorithm appreciation: People prefer algorithmic to human judgment. Organ Behav Hum Decis Process 151:90–103. https://doi.org/10.1016/j.obhdp.2018.12.005
52. Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems, 30*. https://papers.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html
53. Lyell D, Coiera E (2017) Automation bias and verification complexity: A systematic review. J Am Med Inform Assoc 24(2):423–431. https://doi.org/10.1093/jamia/ocw105
54. Martire KA, Kemp RI, Watkins I, Sayle MA, Newell BR (2013) The expression and interpretation of uncertain forensic science evidence: Verbal equivalence, evidence strength, and the weak evidence effect. Law Hum Behav 37(3):197–207. https://doi.org/10.1037/lhb0000027
55. McDowell M, Jacobs P (2017) Meta-analysis of the effect of natural frequencies on Bayesian reasoning. Psychol Bull 143(12):1273–1312. https://doi.org/10.1037/bul0000126
56. Mirsky Y, Lee W (2021) The creation and detection of deepfakes: A survey. ACM-CSUR 54(1) Article 7. https://doi.org/10.1145/3425780
57. Mitchell M, Wu S, Zaldivar A, Barnes P, Vasserman L, Hutchinson B, Spitzer E, Raji ID, Gebru T (2019) Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. https://doi.org/10.1145/3287560.3287596
58. Mnookin JL, Cole SA, Dror IE, Fisher BAJ, Houck MM, Inman K, Kaye DH, Koehler JJ, Langenburg G, Risinger DM, Rudin N, Siegel J, Stoney DA (2011) The need for a research culture in the forensic sciences. UCLA Law Rev 58(3):725–779. https://scholarlycommons.law.northwestern.edu/facultyworkingpapers/26/
59. Mosier KL, Skitka LJ, Heers S, Burdick M (1998) Automation bias: Decision making and performance in high-tech cockpits. Int J Aviat Psychol 8(1):47–63. https://doi.org/10.1207/s15327108ijap0801_3
60. Nightingale SJ, Farid H (2022) AI-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences, 119*(8), e2120481119. https://doi.org/10.1073/pnas.2120481119
61. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hróbjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald S, Moher D (2021) The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. BMJ 372:n71. https://doi.org/10.1136/bmj.n71
62. Parasuraman R, Riley V (1997) Humans and automation: Use, misuse, disuse, abuse. Hum Factors: J Hum Factors Ergon Soc 39(2):230–253. https://doi.org/10.1518/001872097778543886
63. Parasuraman R, Sheridan T, Wickens C (2000) A model for types and levels of human interaction with automation. IEEE Trans Syst Man Cybernetics - Part A: Syst Hum 30(3):286–297. https://doi.org/10.1109/3468.844354
64. Pearson J, Dror IE, Jayes E, Whordley G-R, Mason G, Nightingale S (2026) Examining human reliance on artificial intelligence in decision making. Sci Rep 16:5345. https://doi.org/10.1038/s41598-026-

34983-y

65. Postiglione M, Gortner I, Subrahmanian VS (2026) Human–AI ensembles improve deepfake detection in low-to-medium quality videos [Preprint]. *arXiv*. https://doi.org/10.48550/arXiv.2603.14658
66. Rethlefsen ML, Kirtley S, Waffenschmidt S, Ayala AP, Moher D, Page MJ, Koffel JB, Blunt H, Brigham T, Chang S, Clark J, Conway A, Couban R, de Kock S, Farrah K, Fehrmann P, Foster M, Fowler SA, Glanville J, Young S (2021) PRISMA-S: an extension to the PRISMA Statement for Reporting Literature Searches in *Systematic Reviews*. Syst Reviews 10(1):39. https://doi.org/10.1186/s13643-020-01542-z
67. Ribeiro MT, Singh S, Guestrin C (2016) Why should I trust you? Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. https://doi.org/10.1145/2939672.2939778
68. Risinger DM, Saks MJ, Thompson WC, Rosenthal R (2002) The Daubert/Kumho implications of observer effects in forensic science: Hidden problems of expectation and suggestion. Calif Law Rev 90(1):1. https://doi.org/10.2307/3481305
69. Rössler A, Cozzolino D, Verdoliva L, Riess C, Thies J, Niessner M (2019) FaceForensics++: Learning to detect manipulated facial images. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 1–11. https://doi.org/10.1109/iccv.2019.00009
70. Roth A (2016) Trial by machine. Georget Law J 104(5):1245–1306. https://lawcat.berkeley.edu/record/1127527
71. Roth A (2017) Machine testimony. Yale Law J 126(7):1972–2053. https://yalelawjournal.org/article/machine-testimony
72. Roux C, Crispino F, Ribaux O (2012) From forensics to forensic science. Curr Issues Criminal Justice 24(1):7–24. https://doi.org/10.1080/10345329.2012.12035941
73. Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1(5):206–215. https://doi.org/10.1038/s42256-019-0048-x
74. Saks MJ, Faigman DL (2008) Failed forensics: How forensic science lost its way and how it might yet find it. Annual Rev Law Social Sci 4(1):149–171. https://doi.org/10.1146/annurev.lawsocsci.4.110707.172303
75. Saks MJ, Koehler JJ (2005) The coming paradigm shift in forensic identification science. Science 309(5736):892–895. https://doi.org/10.1126/science.1111565
76. Shea BJ, Reeves BC, Wells G, Thuku M, Hamel C, Moran J, Moher D, Tugwell P, Welch V, Kristjansson E, Henry DA (2017) AMSTAR 2: A critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. BMJ j4008. https://doi.org/10.1136/bmj.j4008
77. Simonsohn U, Nelson LD, Simmons JP (2014) P-curve: A key to the file-drawer. J Exp Psychol Gen 143(2):534–547. https://doi.org/10.1037/a0033242

78. Skitka LJ, Mosier KL, Burdick M (1999) Does automation bias decision-making? Int J Hum Comput Stud 51(5):991–1006. https://doi.org/10.1006/ijhc.1999.0252
79. Somoray K, Miller DJ, Holmes M (2025) Human Performance in Deepfake Detection: A Systematic Review. Hum Behav Emerg Technol 2025(1):1833228. https://doi.org/10.1155/hbe2/1833228
80. Sterne JAC, Savović J, Page MJ, Elbers RG, Blencowe NS, Boutron I, Cates CJ, Cheng H-Y, Corbett MS, Eldridge SM, Emberson JR, Hernán MA, Hopewell S, Hróbjartsson A, Junqueira DR, Jüni P, Kirkham JJ, Lasserson T, Li T, Higgins JPT (2019) RoB 2: A revised tool for assessing risk of bias in randomised trials. BMJ 366:l4898. https://doi.org/10.1136/bmj.l4898
81. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, Henry D, Altman DG, Ansari MT, Boutron I, Carpenter JR, Chan A-W, Churchill R, Deeks JJ, Hróbjartsson A, Kirkham J, Jüni P, Loke YK, Pigott TD, Higgins JP (2016) ROBINS-I: A tool for assessing risk of bias in non-randomised studies of interventions. BMJ 355:i4919. https://doi.org/10.1136/bmj.i4919
82. Suresh H, Guttag J (2021) A framework for understanding sources of harm throughout the machine learning life cycle. *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–9. https://doi.org/10.1145/3465416.3483305
83. Tariq S, Woo SS, Singh P, Irmalasari I, Gupta S, Gupta D (2025) From Prediction to Explanation: Multimodal, Explainable, and Interactive Deepfake Detection Framework for Non-Expert Users. *Proceedings of the 33rd ACM International Conference on Multimedia*, *11716*–11725. https://doi.org/10.1145/3746027.3755786
84. Thompson WC (2009) Painting the target around the matching profile: The texas sharpshooter fallacy in forensic DNA interpretation. Law Probab Risk 8(3):257–276. https://doi.org/10.1093/lpr/mgp013
85. Thompson WC, Newman EJ (2015) Lay understanding of forensic statistics: Evaluation of random match probabilities, likelihood ratios, and verbal equivalents. Law Hum Behav 39(4):332–349. https://doi.org/10.1037/lhb0000134
86. Thompson WC, Schumann EL (1987) Interpretation of statistical evidence in criminal trials: The prosecutor's fallacy and the defense attorney's fallacy. Law Hum Behav 11(3):167–187. https://doi.org/10.1007/bf01044641
87. Tolosana R, Vera-Rodriguez R, Fierrez J, Morales A, Ortega-Garcia J (2020) Deepfakes and beyond: A survey of face manipulation and fake detection. Inform Fusion 64:131–148. https://doi.org/10.1016/j.inffus.2020.06.014
88. Vaccari C, Chadwick A (2020) Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. Social Media + Soc 6(1):2056305120903408. https://doi.org/10.1177/2056305120903408
89. van der Bles AM, van der Linden S, Freeman ALJ, Spiegelhalter DJ (2020) The effects of communicating uncertainty on public trust in facts and numbers. *Proceedings of the National Academy of Sciences, 117*(14), 7672–7683. https://doi.org/10.1073/pnas.1913678117

90. Verdoliva L (2020) Media forensics and DeepFakes: An overview. IEEE J Selec Topics Signal Process 14(5):910–932. https://doi.org/10.1109/jstsp.2020.3002101
91. Wachter S, Mittelstadt B, Russell C (2018) Counterfactual explanations without opening the black box: Automated decisions and the GDPR. Harv J Law Technol 31(2):841–887. https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf
92. Wang S-Y, Wang O, Zhang R, Owens A, Efros AA (2020) CNN-generated images are surprisingly easy to spot… for now. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8692–8701. https://doi.org/10.1109/cvpr42600.2020.00872
93. Westerlund M (2019) The emergence of Deepfake technology: A review. Technol Innov Manage Rev 9(11):39–52. https://doi.org/10.22215/timreview/1282
94. Westfall J, Judd CM, Kenny DA (2015) Replicating studies in which samples of participants respond to samples of stimuli. Perspect Psychol Sci 10(3):390–399. https://doi.org/10.1177/1745691614564879
95. Wexler R (2018) Life, liberty, and trade secrets: Intellectual property in the criminal justice system. Stanford Law Rev 70:1343–1429. https://www.stanfordlawreview.org/print/article/life-liberty-and-trade-secrets/
96. Whiting P, Savović J, Higgins JP, Caldwell DM, Reeves BC, Shea B, Davies P, Kleijnen J, Churchill R (2016) ROBIS: A new tool to assess risk of bias in systematic reviews was developed. J Clin Epidemiol 69:225–234. https://doi.org/10.1016/j.jclinepi.2015.06.005
97. Yarkoni T (2022) The generalizability crisis. Behav Brain Sci 45:e1. https://doi.org/10.1017/s0140525x20001685
98. Zhang M, Kouadri S, Wong P, McJury M, Bharati N, Bandara A (2026) Understanding the role of visual explanations in human–AI collaborations in deepfake image detection [Workshop paper]. https://doi.org/10.5281/zenodo.19700249. HCXAI 2026. Zenodo
99. Zhang Y, Liao QV, Bellamy RKE (2020) Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 295–305. https://doi.org/10.1145/3351095.3372852
100. Zhou Y, Chen E, Pisipati M, Xiong A, Rajtmajer S (2025) Effect of AI Performance, Risk Perception, and Trust on Human Dependence in Deepfake Detection AI System. *Proceedings of the ACM on Human-Computer Interaction*, *9*(7), 1–24. https://doi.org/10.1145/3757588

Figures

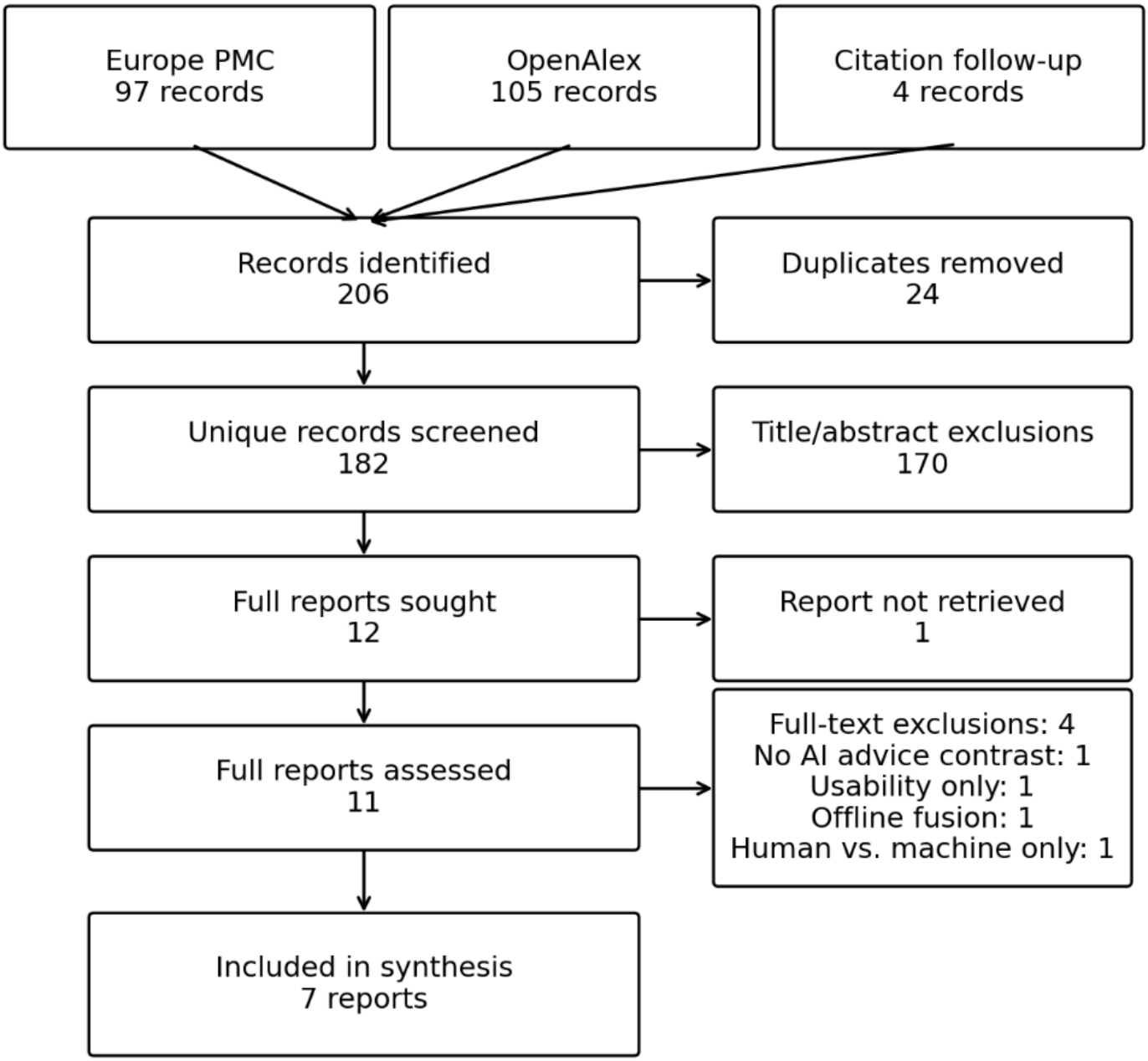


Figure 1

Report selection flow

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Appendixs.docx