AI & Future Society (AFS)



When AI Weakens Evidence Traceability: Ethical Challenges for Credibility and Responsibility in Digital Forensics

Xuzhuo Zhang[a, ∗]

[a]*Faculty of Computer Science and Information Technology, Universiti Malaya, Kuala Lumpur, Malaysia*

DOI: https://doi.org/10.63802/afs.V2.I1.208

ARTICLE INFO

Keywords:
Artificial Intelligence
Digital Forensics
Evidence Traceability
Algorithmic Accountability
AI Governance

ABSTRACT

Artificial intelligence (AI) is increasingly used in digital forensic workflows, offering efficiency benefits but raising critical ethical concerns. This paper examines the weakening of evidence traceability as a fundamental challenge, arguing that reduced traceability threatens evidence credibility and severely erodes responsibility attribution. Unlike conventional technical errors, traceability breakdowns generate plausible analytic statements lacking explicit grounding in identifiable forensic artifacts. Through a normative ethical analysis, the Traceability-Responsibility Convergence framework is introduced to demonstrate how these technical breakdowns are exacerbated by structural accountability gaps and human cognitive biases, such as automation bias and institutional normalisation. The paper concludes that technical validation alone is insufficient. Responsible AI use requires clear human–AI role definition, auditable procedures, and governance frameworks that assign accountability across the AI lifecycle to maintain evidentiary legitimacy.

1. Introduction

Artificial intelligence (AI), including machine learning (ML) and large language models (LLMs), is increasingly integrated into digital forensics to manage the scale and diversity of contemporary evidentiary data, such as communication records, multimedia files, and device artifacts. Prior research in digital forensic literature documents the use of ML techniques to support or automate forensic tasks including triage, prioritisation, and artifact discovery, reflecting the volume of data encountered in investigations (Marturana et al., 2014; Ngejane et al., 2021; Dunsin et al., 2024; Khan et al., 2024).

The transformative impact of AI adoption is evident across diverse sectors, ranging from market differentiation strategies in international education (Hu, 2024) to sustainability-oriented innovation in manufacturing SMEs (Xie et al., 2025; Hu et al., 2026). However, unlike these commercial domains, the integration of AI into digital forensics introduces unique high-stakes challenges regarding judicial liberty and evidence admissibility.

As LLMs enter forensic workflows—for example, in artifact summarisation, query assistance, and report drafting—a specific risk becomes relevant: the weakening of evidence traceability, whereby analytic outputs are no longer clearly linked to specific source artifacts or reproducible analytic steps. Existing research identifies structural reliability limitations in LLM-based systems and examines methods aimed at improving output grounding and control (Huang et al., 2025). Studies in other high-stakes domains, including legal research, show that such systems can produce unreliable outputs, creating risks of over-reliance and downstream error (Magesh et al., 2025). In digital forensics, where conclusions must be traceable to artifacts and methods must withstand examination, AI-mediated outputs with weakened traceability can weaken evidence credibility and complicate responsibility attribution when practitioners rely on AI outputs without defined accountability.

Although research on AI and ML in digital forensics has expanded, existing surveys and applied studies primarily focus on technical capability, performance, and efficiency. They do not offer a systematic ethical analysis of how reduced evidence traceability affects evidentiary credibility and responsibility attribution in human–AI forensic workflows. Recent work acknowledges concerns such as hallucination and transparency but does not examine these issues in relation to forensic epistemic

∗ Corresponding author. e-mail: 22078408@siswa.um.edu.my

© **The Author(s).** This article is published in AI & Future Society under the terms of the **Creative Commons Attribution 4.0 International License (CC BY 4.0)**, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

requirements and accountability structures (Dunsin et al., 2024; Sharma et al., 2025; Studiawan et al., 2025). Research on trustworthy AI emphasises that accountability and governance must be treated as foundational requirements, particularly where AI outputs influence decisions with legal consequences (Liu, 2025).

To address this gap, this paper analyses evidence traceability in AI-assisted digital forensics through two connected ethical dimensions: evidence credibility and responsibility attribution. By examining how hallucinations can enter forensic reasoning and reporting, and how responsibility can become distributed across developers, institutions, and practitioners, the paper aims to clarify the ethical implications of AI-assisted forensics and to inform governance and practice measures consistent with the requirement that forensic claims remain grounded in verifiable artifacts.

Finally, the paper proposes a conceptual framework illustrating how technical opacity, mediated by human and institutional factors, directly leads to the erosion of evidentiary credibility.

2. Contribution and Related Works

Existing research on the integration of machine learning and large language models in forensic tasks focuses primarily on capability, performance, and efficiency. While recent studies acknowledge risks like hallucination, they treat these issues as technical glitches rather than fundamental epistemic and accountability failures.

This paper fills this gap by analysing AI integration through a normative ethical lens, specifically connecting technical opacity to forensic epistemic requirements and accountability structures. Table 1 contrasts this analysis with existing scholarship, illustrating how the Traceability-Responsibility Convergence framework advances the current discourse.

Table 1. Comparative Analysis of AI Governance and Digital Forensics Scholarship

Prior Work	Primary Focus	Overlap with This Paper	What This Paper Adds
Dunsin et al., 2024; Sharma et al., 2025; Studiawan et al., 2025	**Technical capabilities and efficiency of AI/LLMs in forensics.**	**Acknowledges hallucination and transparency risks.**	**Shifts focus from technical metrics to epistemic requirements and accountability structures.**
Floridi et al., 2019; Mittelstadt et al., 2016; Kroll et al. 2017	**Broad algorithmic accountability and AI governance.**	**Emphasises institutional responsibility over technical blame.**	**Applies these broad frameworks specifically to the stringent traceability requirements of digital evidence.**
Parasuraman & Riley 1997; Elish, 2025	**Human-automation interaction and automation bias.**	**Recognises how operators defer to systems and act as "moral crumple zones".**	**Contextualises cognitive biases within the adversarial, high-stakes environment of forensic reporting.**

3. Methodology

3.1. Research Approach and Corpus Selection

This study employs a normative ethical analysis to examine the implications of AI integration in digital forensic workflows. The research methodology follows a three-step qualitative approach:

- Workflow Analysis: First, the study identifies critical integration points of AI (including Machine Learning and Large Language Models) in forensic practice, specifically focusing on triage, artefact summarisation, and report generation.
- Epistemic Evaluation: Second, these integration points are evaluated against established forensic epistemic norms, specifically the requirement for evidence traceability and the legal standards for admissibility (e.g., Daubert standards).
- Ethical Synthesis: Third, the study synthesises the technical risks of 'traceability gaps' (such as hallucination and opacity) with the ethical frameworks of accountability and responsibility distribution.

Data for this analysis is drawn from a comprehensive review of three distinct corpora: technical literature on AI failure modes (e.g., hallucination in LLMs), legal statutes and case law regarding digital evidence admissibility, and existing AI governance frameworks (e.g., NIST AI RMF, OECD Principles).

3.2. Conceptual Framework: The Traceability-Responsibility Cycle

To structure the analysis and address the interconnection between technical limitations and ethical outcomes, this paper proposes a conceptual framework: the Traceability-Responsibility Convergence (see Fig.1).

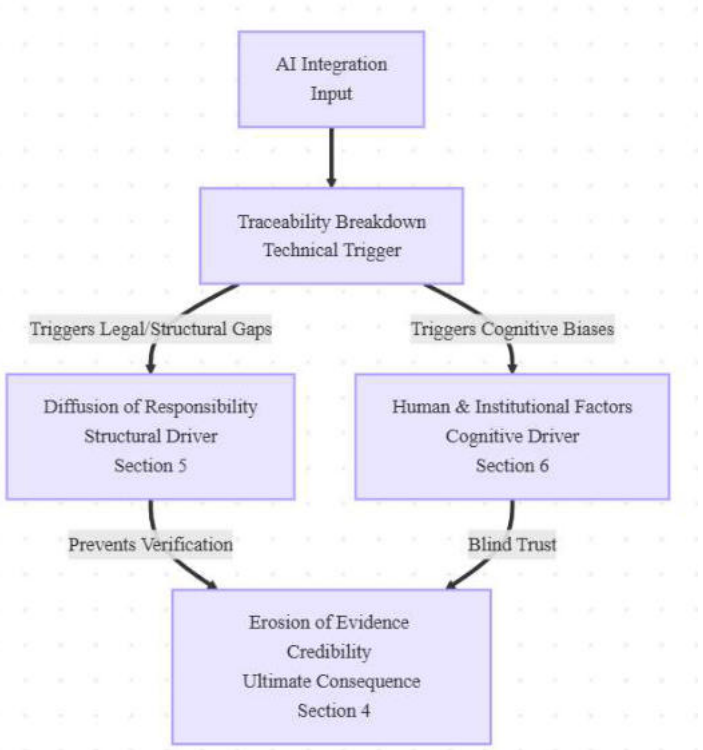


Fig. 1. The Traceability-Responsibility Convergence Framework

As depicted in Figure 1, the technical trigger of a traceability breakdown exerts downward pressure along two distinct vectors. The left directional arrow illustrates how technical opacity triggers structural diffusion of responsibility (Section 5), creating legal accountability gaps. The right directional arrow demonstrates how the same technical failure triggers cognitive human and institutional factors (Section 6), such as automation bias. Ultimately, both vectors converge, resulting in the systemic erosion of evidence credibility. The framework posits that the Traceability Breakdown (the technical trigger) does not act in isolation. Instead, it exacerbates two structural drivers:

1. Human and Institutional Factors (Section 6): Cognitive phenomena such as automation bias and the institutional normalisation of opaque systems.
2. Diffusion of Responsibility (Section 5): A structural gap where accountability is dispersed between vendors, algorithms, and practitioners, creating an environment of 'verification negligence'.

As illustrated in the framework, these factors converge to produce the ultimate forensic failure: a critical Erosion of Evidence Credibility (Section 4). This model demonstrates that credibility is not merely a technical attribute but a systemic outcome dependent on clear responsibility attribution.

4. Evidence Credibility and Evidence Traceability

4.1. Conceptualising Credibility and Admissibility

It is critical to establish a definitive boundary between epistemic credibility and legal admissibility. Credibility functions as an epistemic requirement. It ensures that forensic claims are scientifically grounded in observable artefacts and reproducible methods (Kent et al., 2006). Conversely, admissibility represents a legal threshold, such as the Daubert standard (Supreme Court of the United States, 1993), which governs whether evidence is legally permitted in adversarial proceedings. While admissibility standards evaluate the testability of methods, evidentiary credibility must be secured fundamentally before legal admissibility can even be contested (Casey, 2009). Speculative inference and undocumented reasoning inherently weaken the standing of forensic findings in adversarial proceedings (Howell, 2004). Consequently, a traceability breakdown destroys this epistemic foundation, rendering the subsequent legal assessment moot.

4.2. Operationalising Traceability Breakdown

Reduced evidence traceability refers to situations in which analytic outputs are coherent in form but lack a clear, inspectable linkage to specific source artifacts or reproducible analytic steps.

Building upon the normative requirement for credibility, it is necessary to operationalise when a failure occurs. A traceability breakdown is explicitly defined in this framework as occurring when analytic outputs exhibit one or more of the following conditions:

1. Absence of Artefact Linkage: Generated statements or summaries cannot be directly mapped to specific verifiable source artefacts, for example when a system outputs a communication summary without retaining pointers to the raw data logs.
2. Non-Reproducible Steps: The intermediate analytic processes executed by the AI model to reach a conclusion cannot be logically or mathematically reconstructed by a human examiner.
3. Unsubstantiated Generation: The system introduces events, relationships, or conclusions that possess no grounding in the preserved digital evidence, thereby generating plausible but fabricated statements.

This specific phenomenon is widely recognised in technical discourse as hallucination. Research in natural language generation identifies hallucination as a recurring outcome of neural models, including in tasks such as abstractive summarisation where generated statements diverge from input material (Maynez et al., 2020). Survey research further shows that hallucination arises from model objectives and training processes rather than isolated system faults (Ji et al., 2023).

Within digital forensics, hallucination creates risk because AI-generated outputs may be treated as evidence-based statements. When generative systems are used to summarise logs, construct timelines, or assist in report drafting, hallucinated events or relations may be introduced without reference to any digital artifact. When such outputs are incorporated into forensic reasoning or documentation, the distinction between artifact-derived findings and model-generated inference becomes unclear.

This risk is reinforced by automation bias. Studies of human interaction with automated systems show that users may rely on system outputs even when aware of limitations, particularly in complex or high-stakes contexts (Parasuraman & Riley, 1997). In forensic practice, this tendency can lead practitioners to accept AI-generated statements without independent verification, allowing unsupported content to enter evidentiary accounts and weaken trust in forensic conclusions.

4.3. Limits of Technical Validation

One response to traceability risk involves technical validation methods such as confidence scoring, constrained output generation, or explainable AI techniques. Research indicates, however, that such methods do not ensure evidentiary credibility in decision contexts involving legal consequence. Explanatory approaches often provide post hoc accounts that do not correspond to a model's internal processes, which can obscure rather than clarify the basis of an output (Arrieta et al., 2020).

Rudin argues that reliance on explanations for black-box systems is unsuitable in contexts where errors carry consequence, as explanations may give the appearance of control without enabling verification (Rudin, 2019). More broadly, technical validation methods are designed to improve performance measures rather than to support traceability between outputs and source evidence. In digital forensics, the relevant requirement is not prediction but the ability to verify claims against preserved artifacts using established methods (Monasterio Astobiza, 2019).

For these reasons, technical safeguards cannot replace procedural and ethical commitments to evidence credibility. AI-generated outputs must remain subordinate to artifact-based analysis and human judgment that can be examined and attributed. Without such constraints, hallucination risks undermining trust in digital forensic evidence.

4.4. Illustrative Scenarios of Traceability Failure

To demonstrate the practical implications of reduced traceability, two documented real-world failures are presented. The first illustrates a generative hallucination within a legal workflow, while the second details an opaque triage error in a digital multimedia investigation.

Scenario A: The Generative Hallucination Gap (Mata v. Avianca) The first manifestation of traceability failure occurs when generative models produce unsubstantiated content. A documented judicial failure illustrating this phenomenon is the 2023 case of Mata v. Avianca Airlines (Mata v. Avianca, Inc., 2023). Legal practitioners submitted a filing synthesised by a large language model, which generated plausible but entirely fabricated judicial precedents. The system produced non-existent case law lacking any linkage to actual judicial artefacts. Such generative errors, where models produce unsubstantiated content diverging from factual inputs, are well-documented technical phenomena (Huang et al., 2025; Maynez et al., 2020). When the court demanded verification, the traceability breakdown became critically apparent because the outputs could not be traced to valid sources. The incident highlighted a severe diffusion of responsibility, as the practitioners initially deflected blame onto the opacity of the automated tool, demonstrating how the absence of verifiable source grounding directly compromises evidentiary credibility.

Scenario B: The Opaque Triage (The Robert Williams Case) The second manifestation occurs in discriminative machine learning, specifically during automated evidence triage. In 2020, Robert Williams became the subject of a wrongful arrest resulting directly from a flawed facial recognition algorithm (Williams v. City of Detroit, 2024). Investigators utilised an automated system to triage poor-quality surveillance footage, which incorrectly output Williams's driver's licence as a match. Exhibiting severe automation bias (Parasuraman & Riley, 1997), the investigators accepted this opaque output without critically verifying the algorithmic conclusion against the original source artefact. They subsequently used the machine-generated lead to secure an arrest warrant. This case illustrates an opaque triage failure wherein human practitioners defer to a black-box technical system (Rudin, 2019), thereby abandoning the requirement for independent artefact verification and displacing accountability onto the software.

5. Responsibility Attribution in AI-Assisted Forensics

Following the technical breakdown of traceability, a profound structural vulnerability emerges within the forensic workflow. The introduction of AI distributes responsibility across developers, tool providers, and institutions, effectively creating an accountability void. When errors are attributed to model design or training data without a defined legal process for assigning liability (Tiwari et al., n.d.), no single actor remains accountable for the forensic outcomes.

This responsibility diffusion can produce accountability gaps in which no actor can be identified as responsible for erroneous or misleading outputs. Such gaps are associated with socio-technical systems where system behaviour cannot be fully explained or controlled by either system designers or system users (Resnik & Hosseini, 2025). Legal scholarship addressing AI-generated forensic evidence notes difficulties related to reproducibility and validation of AI outputs, which complicate judicial assessment of responsibility for forensic conclusions (Singh & Devi, 2025). Ethical analyses of AI systems emphasise that accountability does not arise from technical design alone but requires governance structures and professional standards that specify responsibility allocation (Coltri et al., 2025). Because forensic conclusions significantly influence legal outcomes and individual rights (Marsico & Amigo, 2025), a process that produces claims without identifiable human accountability undermines procedural justice (Aleke & Trigui, 2025). Without such structures, AI-assisted forensic practice may produce conclusions for which no human agent can be held answerable, weakening institutional trust and the ethical basis of forensic decision-making. Therefore, diffusion of responsibility is not merely an administrative concern but a direct threat to forensic integrity. When accountability is dispersed, the rigorous verification of artefacts and the transparent documentation essential for technical reporting (Horsman, 2022; SWGDE, 2018), which form the cornerstone of evidence credibility, are neglected. Without a clearly defined agent

responsible for validating AI outputs, the link between evidence and inference is broken, rendering the resulting forensic claims legally inadmissible.

6. Human and Institutional Implications

While Section 5 detailed the structural accountability gaps, it is equally critical to examine the behavioural and cognitive effects on practitioners and the environments in which they operate.

In the context of this analysis, the term "forensic institution" refers to the ecosystem of organisations responsible for the processing and adjudication of digital evidence. This specifically includes:

- Public Crime Laboratories and Digital Forensic Units (DFUs): The operational bodies within law enforcement agencies where evidence is processed.
- Private Forensic Firms: Entities contracted to perform analysis in civil or criminal matters.
- Judicial Adjudication Bodies: The courts and legal tribunals that evaluate the admissibility and weight of forensic reports.

Within these environments, cognitive phenomena, such as automation bias, compel forensic examiners to defer judgment to automated outputs, severely reducing independent verification against source artefacts (Dzindolet et al., 2002; Parasuraman & Riley, 1997). This reliance alters how practitioners understand analytic processes, as systems present conclusions without exposing the underlying reasoning steps (Endsley, 2017), leading operators to treat the system as the ultimate source of authority (Hefetz, 2025). Concurrently, these institutions undergo a process of normalisation (Elish, 2025), wherein AI-generated outputs are embedded into routine procedures as unquestioned facts rather than interpretations subject to review (Kroll et al., 2017).

This environment presents a significant risk of ethical laundering, in which responsibility for contested outcomes is attributed to technical systems rather than to human or organisational decision-making. Research in AI ethics describes how references to algorithmic processes can be used to deflect responsibility and to frame outcomes as the result of system operation rather than institutional choice (Ananny & Crawford, 2018). Mittelstadt et al. (2016) argue that algorithmic systems can support this deflection by presenting outcomes as determined by technical processes rather than by normative decisions. In digital forensics, ethical laundering may occur when errors or disputed conclusions are attributed to AI tools, datasets, or vendors instead of to decisions about system selection, deployment, and oversight. This form of responsibility displacement weakens the connection between evidence, judgment, and accountability within forensic institutions. Work on AI governance emphasises that accountability requires institutional structures that assign responsibility for system use and outcomes rather than allowing responsibility to be attributed to technology itself (). Without such structures, AI-assisted forensics may operate in a manner where responsibility is acknowledged in principle but not assigned in practice.

7. Toward Responsible Governance of AI in Digital Forensics

7.1. Human–AI Role Delineation

Governance of AI in digital forensics requires delineation of roles between human practitioners and AI systems. Governance frameworks state that AI systems should function within processes that preserve human oversight and accountability in domains where decisions may affect legal status or rights. The OECD AI Principles specify that humans must retain oversight and responsibility for decisions informed by AI systems (OECD, 2019).

Within digital forensics, this position requires that AI systems be used as analytic tools rather than as decision-making agents. Forensic conclusions must remain attributable to human examiners who can explain and defend their reasoning when conclusions are examined. When governance structures fail to define the limits of AI involvement, responsibility may shift from human actors to technical systems, which complicates accountability within forensic practice.

The NIST AI Risk Management Framework identifies human oversight and role definition as elements of AI governance across system development and use (National Institute of Standards and Technology, 2023). Applied to digital forensics, this requires procedures that specify how AI outputs may inform forensic reasoning and when conclusions must be formed through direct human judgment based on artifacts and methods.

7.2. Procedural and Organisational Safeguards

Role delineation alone does not ensure accountability. Governance of AI in forensic workflows also requires procedures that structure how AI systems are used and reviewed. Research on algorithmic governance emphasises that accountability depends on institutional processes such as documentation, review, and mechanisms for contestation rather than on system transparency alone (Mittelstadt et al., 2016).

In digital forensics, such procedures include documentation of AI tool selection, system versioning, input data, and interpretation of outputs. These practices support audit and review of AI-assisted analysis and allow conclusions to be examined in relation to source artifacts. Without procedural embedding, AI outputs may be incorporated into forensic workflows without mechanisms for review or correction (Ananny & Crawford, 2018).

Governance frameworks addressing algorithmic accountability further indicate that responsibility spans multiple actors, including system developers, deploying organisations, and system users. Kroll et al. describe governance models that combine constraints during system design with processes for review and response after deployment (Kroll et al., 2017). In digital forensics, this approach reflects the distribution of responsibility across tool design, institutional deployment, and practitioner use.

7.3. Ethics-by-Design and Institutional Accountability

Governance of AI in digital forensics also requires integration of ethical considerations into system development and institutional practice. Ethics-by-design approaches argue that responsibility and accountability must be incorporated throughout the AI lifecycle rather than addressed only after system deployment (Dignum, 2019).

Research on AI governance further holds that institutions deploying AI systems in domains involving legal processes retain responsibility for outcomes produced through system use. Accountability frameworks emphasise that responsibility must not be transferred to technical systems, even when system behavior is complex (Floridi et al., 2019). In digital forensics, this requires institutions to accept responsibility for conclusions derived from AI-assisted workflows.

Governance frameworks for AI-assisted forensics should therefore combine role definition, procedural controls, and institutional responsibility into a unified approach. Such governance structures support the maintenance of evidence credibility and responsibility attribution within forensic practice

and align with broader approaches to AI governance across legal and institutional contexts.

7.4. Limitations and Future Work

The analysis presented relies on a normative ethical framework and lacks empirical validation. The arguments concerning automation bias, ethical laundering, and institutional normalisation are extrapolated from broader human computer interaction studies and applied to forensic workflows. Furthermore, the Traceability-Responsibility Convergence framework is primarily conceptualised for high-stakes adversarial legal contexts. Future research must incorporate empirical methodologies, such as observational studies within active digital forensic units and controlled trials with examiners utilising large language models, to accurately quantify the extent of automation bias and responsibility displacement in daily forensic practice.

8. Conclusion

The use of artificial intelligence in digital forensics introduces changes to forensic workflows that raise ethical questions. This paper has argued that weakened evidence traceability functions as a structural risk within digital forensic practice because it affects two requirements of forensic work: evidence credibility and responsibility attribution. In a field where conclusions must be grounded in artifacts and methods that permit examination, the production of unsupported outputs disrupts the connection between evidence and inference.

The analysis of evidence credibility shows that AI-mediated outputs with weakened traceability can obscure the distinction between findings derived from artifacts and statements generated by models. This effect conflicts with forensic norms that require traceability, reproduction, and examination of evidentiary claims. Technical validation methods do not resolve this conflict because they do not ensure that claims remain verifiable through reference to artifacts and accepted forensic procedures.

The analysis of responsibility attribution shows that AI-assisted workflows can distribute responsibility across analysts, systems, developers, and institutions. When responsibility is distributed without defined roles, accountability gaps may arise. In forensic contexts, such gaps are incompatible with the requirement that conclusions be attributable to actors who can provide justification and respond to challenges.

The paper has also examined effects on practitioners and institutions, including changes in how analytic tasks are performed, how system outputs are incorporated into organisational processes, and how responsibility may be displaced from human decision-making to technical systems. These effects can alter the role of AI outputs within forensic practice and affect how evidentiary claims are treated.

In response to these issues, the paper has argued for a governance approach rather than reliance on technical measures alone. Governance of AI in digital forensics requires definition of human and system roles, procedures that support review and contestation of AI-assisted analysis, and institutional structures that assign responsibility across the AI lifecycle. These elements are required to preserve evidence credibility and responsibility attribution within forensic practice.

The acceptability of AI use in digital forensics therefore depends not on the ability of systems to produce outputs, but on whether forensic institutions maintain conditions under which credibility, responsibility, and accountability are sustained. Meeting these conditions is necessary for the use of AI within legal and institutional decision-making processes.

Author Contribution

The author solely conceived the research idea, conducted the literature review, developed the theoretical framework, and completed the writing of the manuscript. All analyses, interpretations, and revisions were carried out independently by the author. The author has read and approved the final version of the manuscript.

Compliance with Ethical Standards

This research does not involve human participants, animals, or sensitive personal data.

The study is based solely on conceptual analysis and publicly available academic literature.

Therefore, formal ethical approval was not required.

Funding

The author received no financial support for the research, authorship, or publication of this article.

Competing Interest declaration

The author declares that there are no competing interests regarding the publication of this paper.

Acknowledgments

The author would like to thank the anonymous reviewers and the editorial board for their constructive comments and suggestions, which greatly improved the quality of this manuscript.

References

1. Aleke, N. T., & Trigui, M. (2025). Legal and ethical challenges in digital forensics investigations. In *Digital Forensics in the Age of AI* (pp. 147–176). IGI Global.
2. Ananny, M., & Crawford, K. (2018). Seeing without knowing: Limitations of the transparency ideal and its application to algorithmic accountability. *New Media & Society*, *20*(3), 973–989.
3. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115.
4. Casey, E. (2009). Handbook of digital forensics and investigation. Academic Press.
5. Coltri, M. A., Uys, W. R., & Joubert, K. (2025). Investing AI ethics in forensic investigations: Development, policies, and best practices. *Athens Journal of Sciences*, 355.
6. Dignum, V. (2019). Responsible artificial intelligence: How to develop and use AI in a responsible way (Vol. 2156). Springer.
7. Dunsin, D., Ghanem, M. C., Ouazzane, K., & Vassilev, V. (2024). A comprehensive analysis of the role of artificial intelligence and machine learning in modern digital forensics and incident response. *Forensic Science International: Digital Investigation*, *48*, 301675.
8. Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors*, *44*(1), 79–94.
9. Elish, M. C. (2025). Moral crumple zones: Cautionary tales in human–robot interaction. In *Robot Law: Volume II* (pp. 83–105). Edward Elgar Publishing.
10. Endsley, M. R. (2017). From here to autonomy: Lessons learned from human–automation research. *Human Factors*, *59*(1), 5–27.
11. Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., ... & Vayena, E. (2018). AI4People—An ethical

framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines*, *28*(4), 689–707.
12. Hefetz, I. (2025). Evaluating bias in forensic evidence: From expert analysis to AI-based decision tools. *Forensic Science International: Synergy*, *11*, 100645.
13. Horsman, G. (2022). Technical reporting in digital forensics. *Journal of Forensic Sciences*, *67*(6), 2458–2468.
14. Howell, B. A. (2004). Ambiguities in US law for investigators. *Digital Investigation*, *1*(2), 106–111.
15. Hu, Y. (2024). The market strategy of the international education industry: A differentiation perspective in research. *SHS Web of Conferences*, *187*, 04011. EDP Sciences.
16. Hu, Y., Jusoh, S., Abd Razak, M. F., Yusof, A. R. B. M., Xie, Q., Song, C., & Zong, X. (2026). Understanding the digital transformation of SMEs: A systematic review from a dynamic capabilities perspective across business, technology, and societal domains. *Global Review of Humanities, Arts, and Society*, *2*(1), 1–17.
17. Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... & Liu, T. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, *43*(2), 1–55.
18. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1–38.
19. Kent, K., Chevalier, S., Grance, T., & Dang, H. (2006). *Guide to integrating forensic techniques into incident response* (NIST Special Publication 800-86). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-86
20. Khan, A. A., Zhang, X., Hajjej, F., Yang, J., Ku, C. S., & Por, L. Y. (2024). ASMF: Ambient social media forensics chain of custody with an intelligent digital investigation process using federated learning. *Heliyon*, *10*(1).
21. Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. *University of Pennsylvania Law Review*, *165*(3), 633–705.
22. Liu, E. (2025). Building trustworthy AI: Transparency, fairness, and governance in the digital age. *AI & Future Society*, *1*(1), 3–5.
23. Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., & Ho, D. E. (2025). Hallucination-free? Assessing the reliability of leading AI legal research tools. *Journal of Empirical Legal Studies*, *22*(2), 216–242.
24. Marsico, F., & Amigo, M. (2025). Ethical and security challenges in AI for forensic genetics: From bias to adversarial attacks. *Forensic Science International: Genetics*, *76*, 103225.
25. Marturana, F., & Tacconi, S. (2013). A machine learning-based triage methodology for automated categorization of digital media. *Digital Investigation*, *10*(2), 193–204.
26. Mata v. Avianca, Inc., 22-cv-1461 (S.D.N.Y. 2023).
27. Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. *arXiv preprint arXiv:2005.00661*.
28. Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society*, *3*(2).
29. Monasterio Astobiza, A., Toboso, M., Aparicio, M., Ausín, T., López, D., Morte, R., & Pons, J. L. (2019). Bringing inclusivity to robotics with INBOTS. *Nature Machine Intelligence*, *1*(4), 164.
30. National Institute of Standards and Technology. (2023). *AI risk management framework (AI RMF 1.0)*. https://doi.org/10.6028/NIST.AI.100-1
31. Ngejane, C. H., Eloff, J. H., Sefara, T. J., & Marivate, V. N. (2021). Digital forensics supported by machine learning for the detection of online sexual predatory chats. *Forensic Science International: Digital Investigation*, *36*, 301109.
32. OECD. (2019). OECD principles on artificial intelligence. OECD Publishing.
33. Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253.
34. Resnik, D. B., & Hosseini, M. (2025). The ethics of using artificial intelligence in scientific research: New guidance needed for a new tool. *AI and Ethics*, *5*(2), 1499–1521.
35. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215.
36. Scientific Working Group on Digital Evidence. (2018). *Requirements for report writing in digital and multimedia forensics* (18-Q-002-1.0). https://www.swgde.org/wp-content/uploads/2023/11/2018-11-20-SWGDE-Requirements-for-Report-Writin.pdf
37. Sharma, B., Ghawaly, J., McCleary, K., Webb, A. M., & Baggili, I. (2025). ForensicLLM: A local large language model for digital forensics. *Forensic Science International: Digital Investigation*, *52*, 301872.
38. Singh, S., & Devi, L. (2025). Reliability and admissibility of AI-generated forensic evidence in criminal trials. *arXiv preprint arXiv:2601.06048*.
39. Studiawan, H., Breitinger, F., & Scanlon, M. (2025). Towards a standardized methodology and dataset for evaluating LLM-based digital forensic timeline analysis. *arXiv preprint arXiv:2505.03100*.
40. Supreme Court of the United States. (1993). *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579.
41. Tiwari, V., Dasari, V. S. R., & Wang, J. (n.d.). The future of artificial intelligence in forensics: Advancements, challenges, and ethical considerations.
42. Williams v. City of Detroit, No. 2:21-cv-10827 (E.D. Mich. 2024).
43. Xie, Q., Hanafiah, M. H., Yang, H., Abdul Hamid, R., & Juhdi, N. H. (2025). Mediating effects of sustainability-oriented innovation on the relationship between organizational creativity and sustainable performance among Chinese manufacturing SMEs. *Multidisciplinary Reviews*, *8*(12), 2025373. https://doi.org/10.31893/multirev.2025373