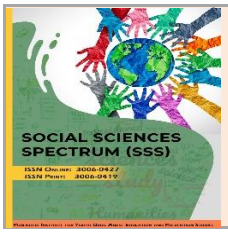

Social Sciences Spectrum

A Double-Blind, Peer-Reviewed, HEC recognized Y-category Research Journal

E-ISSN: 3006-0427 **P-ISSN:** 3006-0419
Volume 04, **Issue** 02, 2025
Web link:https://sss.org.pk/index.php/sss

Check for updates

Ethical Governance of Artificial Intelligence Hallucinations in Legal Practice

Muhammad Khurram Shahzad Warraich
Assistant Director, Research Provincial Assembly of the Punjab Lahore, Punjab, Pakistan
Emai: khurramwarraich882@gmail.com

Hazrat Usman
Advocate High Court, Punjab Bar Council, Punjab, Pakistan
Correspondence: hazratusmanadvocate@gmail.com

Sidra Zakir
LLB Student, Department of Law, Mohi Ud Din Islamic University Nerian Sharif Azad Jammu & Kashmir, Pakistan
Emai: sidrazakir508@gmail.com

Dr. Mohaddas Mehboob
Assistant Professor, Department of Law, Ibadat International University Islamabad, Pakistan
Email: mohaddas.naqvi@law.iiui.edu.pk

Article Information [YY-MM-DD]

Received 2025-03-26 **Accepted** 2025-05-30

Citation (APA):

Warraich, M, K, S., Usman, H., Zakir, S & Mehboob, M. (2025). Ethical Governance of artificial intelligence Hallucinations in legal practice. *Social Sciences Spectrum*, *4*(2), 603-615. https://doi.org/10.71085/sss.04.02.297

Abstract

This paper examines the ethical and legal challenges posed by "hallucinations" in generative-AI tools used for legal drafting—instances where language models fabricate case citations or statutory text with convincing authority. Drawing on a comprehensive review of professional-responsibility rules, civil-liability doctrines, and technical mitigation strategies, the study assesses how existing frameworks address, or fail to prevent, AI-induced errors in attorney filings. Empirical benchmarking data reveal that leading retrieval-augmented models still produce fabricated authorities in up to one-third of complex queries, while sanctions under traditional malpractice and negligence regimes remain retrospective and inconsistent. Comparative analysis of U.S. and EU liability proposals—the AI Liability Directive and the Revised Product Liability Directive—highlights gap in coverage for bespoke legal services. In response, the paper proposes an integrated governance model combining binding bar-association standards (mandatory AI-literacy training, provenance logging, and human-in-the-loop review), statutory safe-harbor provisions granting rebuttable presumptions of compliance, and robust technical protocols. The study concludes by recommending targeted rule-making, pilot programs to evaluate framework efficacy, and incorporation of AI governance curricula in legal education, thereby safeguarding the integrity of legal practice in the AI era.

Keywords: Generative Artificial Intelligence, Hallucinated Legal Citations, Ethical Governance, Retrieval-Augmented Verification, AI Liability Frameworks, Comparative Legal Ethics.

Introduction

Artificial intelligence (AI) has rapidly transformed legal research and practice, offering unprecedented efficiencies in tasks ranging from document review to predictive analytics. Yet this surge in AI adoption has unveiled a critical and underexplored challenge: the phenomenon of "hallucinations," wherein generative models fabricate case citations, statutory provisions, or factual details with high confidence (Stanford HAI, 2023). In June 2025, for example, at least 58 instances of AI-generated "phantom" precedents appeared in attorney filings, prompting courts to question the reliability of AI-assisted briefs (Stanford HAI, 2023). Such inaccuracies not only mislead judicial decision-making but also expose legal practitioners to sanctions and reputational harm, underscoring an urgent need for ethical governance frameworks tailored to legal contexts (The Guardian, 2023; AP News, 2023).

Hallucinations arise primarily from the probabilistic nature of large language models (LLMs) that prioritize fluency and pattern-matching over verifiable truth (Epstein Becker & Green, 2024). In the absence of robust verification protocols, attorneys relying on LLM outputs risk embedding fictional authorities into court documents, as illustrated by a U.S. Southern District of New York sanction in Mata v. Avianca, where the court reprimanded counsel for citing cases that did not exist (Epstein Becker & Green, 2024). Furthermore, comparative analyses reveal similar incidents in the United Kingdom, where the High Court warned that unvetted AI citations threatened the very integrity of legal process (The Guardian, 2023). These judicial responses signal nascent but fragmented efforts to address AI-driven errors under existing professional-responsibility rules.

Current normative instruments—including the American Bar Association's Model Rules of Professional Conduct and the District of Columbia Bar Ethics Opinion 388—mandate competence, Candor, and client confidentiality but offer limited guidance on AI-specific risks (American Bar Association, 2024). Rule 1.1 requires attorneys to maintain "the legal knowledge, skill, thoroughness, and preparation reasonably necessary" for representation, yet does not specify how to assess the reliability of generative tools (American Bar Association, 2024, p. 5). Formal Opinion 512 (2024) advises that lawyers must understand an AI system's capabilities and limitations but stops short of prescribing technical safeguards such as retrieval-augmented verification or human-in-the-loop review processes (American Bar Association, 2024). Similarly, data-protection regimes like the General Data Protection Regulation (GDPR) enshrine principles of transparency and accountability—mandating that automated decisions be explainable under Articles 5, 15, and 22—but do not directly confront the peculiarities of hallucination in legal drafting (European Parliament, 2016; Kharitonova, 2022).

At the supranational level, the European Commission's proposed Artificial Intelligence Act (2021) adopts a risk-based approach, labelling legal-reasoning systems as "high risk" and imposing conformity assessments and post-market monitoring (European Commission, 2021). Yet the Act emphasizes safety and fundamental-rights protection rather than professional-responsibility norms, leaving a regulatory lacuna regarding the accuracy of AI-generated legal citations. High-Level Expert Group guidelines on "Trustworthy AI" call for human oversight, technical robustness, and clear governance structures, but their nonbinding nature limits enforceability (High-Level Expert Group on AI, 2019).

This research paper addresses these gaps by proposing an integrated ethical-governance model to prevent and remediate AI hallucinations in legal practice. First, it systematically reviews documented instances of hallucinated citations and appellate responses, drawing on case studies from the United States, United Kingdom, and European courts (AP News, 2023; Epstein Becker

& Green, 2024). Next, it examines existing professional-responsibility rules, data-protection statutes, and AI-specific regulations to identify overlaps, tensions, and lacunae. The analysis then turns to technical mitigation strategies—such as retrieval-augmented generation, provenance tracking, and human-in-the-loop checkpoints—evaluating their compatibility with legal-ethical obligations (Stanford HAI, 2023).

Drawing on doctrinal and comparative methods, the paper advances a normative framework that combines: (a) mandatory disclosure of AI-assisted drafting, (b) standardized verification protocols codified by bar associations, and (c) enforcement mechanisms aligned with disciplinary procedures. This tripartite approach seeks to uphold the integrity of judicial processes while allowing law firms and courts to benefit from AI's efficiencies. By harmonizing technical safeguards with ethical duties, the framework aspires to transform ad hoc judicial admonitions into coherent governance standards that can be adopted by professional bodies worldwide.

In delineating this model, the paper contributes to scholarly and practical debates at the intersection of AI ethics and legal regulation. It offers actionable recommendations for policymakers, bar associations, and legal educators to integrate AI literacy and ethical training into curricula and continuing-legal-education programs. Ultimately, the study aims to foster a culture of responsibility and transparency in AI-enhanced legal practice—ensuring that technological innovation reinforces, rather than undermines, the rule of law.

Reconciling Professional Responsibility and Technical Safeguards in Generative AI Use

The accelerating incorporation of generative AI tools into legal workflows has exposed a critical ethical fault line: large language models (LLMs) can "hallucinate" authoritative—but non-existent—case citations or statutory provisions, compromising both the reliability of legal documents and the practitioner's professional obligations (Stanford HAI, 2024). Under the American Bar Association's Model Rules of Professional Conduct, lawyers must maintain competence (Rule 1.1) and candour toward tribunals (Rule 3.3), yet these provisions predate generative AI and offer no bespoke mechanisms to verify the factual integrity of algorithm-generated content (American Bar Association, 2024). In Mata v. Avianca, the Southern District of New York sanctioned counsel for relying on ChatGPT-fabricated cases, underscoring that existing malpractice and evidentiary rules can penalize hallucinations only after harm occurs—not prevent them proactively (Mata v. Avianca, 2023). Similarly, more than twenty state and federal jurisdictions have issued standing orders requiring attorneys to disclose or supervise AI use, reflecting a nascent, piecemeal regulatory response rather than a unified professional standard (Stanford HAI, 2024).

Early doctrinal scholarship has noted that these disciplinary and evidentiary rules rest on principles of reasonableness and diligence, yet they lack calibrated benchmarks for AI reliability (Epstein Becker & Green, 2024). The duty of competence implicitly demands that an attorney understand an AI system's capabilities and limitations, but without standardized technical protocols—such as retrieval-augmented verification or provenance tracking—lawyers face an impossible burden of manually vetting every AI-generated proposition (Stanford HAI, 2024). The disconnect between longstanding professional norms and the opaque workings of generative models reveals a significant blind spot: ethical rules clearly demand accuracy, candour, and thoroughness, yet they offer little guidance on the concrete steps lawyers must take when AI is in the drafting process. To bridge this divide, we must embed tangible technical checks into the existing ethical framework—essentially translating broad duties into a series of practical, verifiable stages that ensure every AI-suggested citation is rigorously confirmed before it enters a brief.

One promising solution is retrieval-augmented generation (RAG), which anchors a model's output in an approved collection of legal materials (Stanford HAI, 2024). Under this method, the AI first scours a curated database to surface directly relevant cases or statutes and then weaves those sources into its response. Although RAG significantly curtails the incidence of fabricated authorities, it does not eradicate them: tests run on popular platforms like Lexis+ AI and Westlaw AI-Assisted Research still reveal error rates of around 17 percent and 34 percent, respectively, underscoring that technical safeguards must be paired with human oversight to achieve near-perfect reliability (Stanford HAI, 2024). Contributing factors include retrieval failures—where semantically similar but legally inapposite documents are returned—and generation errors that mis render or fabricate citations, a problem exacerbated in areas of rapidly evolving law (Stanford HAI, 2024).

Beyond RAG, transparency mechanisms drawn from data-protection law offer instructive analogies. The General Data Protection Regulation mandates explainability and provenance for automated decisions (Arts. 5, 15, 22), requiring controllers to provide meaningful information about algorithmic logic and to allow challenge or review (European Parliament, 2016). Kharitonova's comparative study of AI transparency in the EU and Russia argues for codifying obligations to log retrieval sources and to attach confidence scores or provenance metadata to every AI-generated citation (Kharitonova, 2022). Likewise, Buchner's analysis of "Doctor Algorithm" emphasizes that legal AI systems—characterized as "high risk" under the proposed EU AI Act—must satisfy robustness and auditability requirements to ensure human actors can evaluate and correct outputs (Buchner, 2022).

These technical and legal threads converge on the need for integrated normative frameworks—standards that map professional-responsibility duties onto concrete technical protocols. The European Commission's AI Act proposal classifies legal-reasoning systems as high risk, mandating conformity assessments, post-market monitoring, and human oversight (Arts. 5 & 14) (European Commission, 2021). Although comprehensive, the Act's focus on safety and fairness omits guidance on citation accuracy, leaving a regulatory lacuna for professional bodies to fill. The High-Level Expert Group's nonbinding "Trustworthy AI" guidelines prescribe human oversight, technical robustness, and clear governance structures—tenets that, if codified by bar associations, could bridge the gap between algorithmic controls and ethical duties (High-Level Expert Group on AI, 2019). Pistilli, Muñoz Ferrandis, Jernite, and Mitchell (2023). further recommend aligning AI liability principles (as in the forthcoming AI Liability Directive) with professional-responsibility sanctions, so that malpractice claims reflect both technical failures and breaches of ethical norms (Pistilli et al., 2023).

Together, these strands of literature underscore that neither professional rules nor technical safeguards alone suffice ethical governance of AI hallucinations demands a tripartite model combining (a) explicit professional-responsibility rules requiring AI-disclosure and verification, (b) mandatory technical protocols—RAG, provenance logging, and confidence scoring—and (c) enforcement mechanisms aligned with disciplinary and liability frameworks. Only through such an integrated approach can the legal profession harness generative AI's efficiencies without sacrificing the integrity of legal reasoning.

Liability Frameworks for AI-Induced Errors in Legal Submissions

The emergence of generative AI tools in legal practice has not only raised questions of professional ethics but also sparked urgent debates about civil liability when AI-induced errors cause harm. Hallucinations—AI's confident fabrication of case law, statutory text, or factual assertions—have

already led to sanctions and court admonitions (Mata v. Avianca, 2023). Yet beyond disciplinary measures under professional-responsibility rules, litigants who suffer prejudice from such errors must seek redress through malpractice or negligence claims. Traditional legal malpractice doctrine in the United States requires proof of (a) an attorney's duty of care, (b) a breach of that duty, (c) causation linking the breach to the client's harm, and (d) compensable damages (Bashayreh, Tabbara, & Sibai, 2024). In the context of AI hallucinations, courts have signalled that submitting a brief with fabricated authorities can satisfy breach and causation—but only after a post-hoc review reveals prejudice to the client or the administration of justice (Epstein Becker & Green, 2024). This retrospective, fault-based process mirrors common law negligence but leaves a regulatory gap: clients and opposing parties must wait for demonstrable harm before malpractice insurers, bar associations, or courts address algorithmic errors.

Evidence from related domains underscores the limitations of existing fault-based frameworks when applied to AI. In healthcare, for example, clinical-decision support systems (CDSS) have produced diagnostic mistakes, prompting scholars to call for a "legal standard of care" specific to AI tools—one that delineates both clinicians' and vendors' duties to ensure algorithmic reliability (Prictor, 2023; Rowland et al., 2022). Bashayreh et al. (2024) similarly argue that attorneys need a comparable standard tailored to AI-assisted drafting, clarifying when a lawyer's reliance on AI constitutes reasonable care and when it amounts to negligence. Absent such guidance, courts must shoehorn AI errors into broad negligence principles, resulting in inconsistent outcomes (Giannini & Kwik, 2022). Moreover, the battleground over burden of proof in AI-related malpractice claims remains unsettled: Llorca, Charisi, Hamon, Sánchez, and Gómez (2023) demonstrate that existing tort regimes place the evidentiary burden on plaintiffs to prove both the AI's malfunction and the attorney's failure to exercise due care, a daunting task when model architectures and training data are proprietary.

In response to these challenges, the European Union has pioneered specialized liability frameworks aimed at lowering barriers to compensation for AI-related harms. The Proposed Directive on Adaptation of Non-Contractual Civil Liability Rules to Artificial Intelligence (AI Liability Directive, AILD) and the accompanying revision of the Product Liability Directive (PLD) seek to introduce ex-ante obligations for high-risk AI systems and shift certain burdens of proof (European Commission, 2022a, 2022b). Under the AILD, victims need only demonstrate (1) that harm occurred, (2) that the damage was caused by an AI system operating in a high-risk context, and (3) that the defendant failed to meet specified obligations—such as data-governance standards or human-oversight requirements—before liability is presumed (Hacker, 2023). This presumption can be rebutted, but it alleviates plaintiffs' need to prove intricate technical causation. Rodríguez de las Heras Ballell (2023) notes that, although the Revised PLD extends strict liability for defective products to certain AI applications, gaps remain—particularly for professional services like legal drafting that do not fall neatly within product-liability categories.

Scholars critique these EU initiatives for their uneven scope and potential to generate fragmentation across member states. Duffourc and Gerke (2023) warn that medical-AI providers may benefit from the PLD's safe harbour while legal-AI developers—whose tools often integrate into bespoke workflows—could slip through regulatory cracks. The European Law Institute's response to the public consultation highlights tensions between liability deterrence and innovation, arguing that overly broad strict-liability regimes could stifle AI adoption while narrow fault-based exceptions perpetuate victims' burdens (European Law Institute, 2022). Chamberlain's risk-based analysis further suggests that the EU's tiered approach to AI classification, if aligned with liability regimes, can calibrate obligations so that only truly hazardous applications trigger presumptions

of fault or strict liability (Chamberlain, 2021). Yet absent explicit references to professional-practice contexts, these directives risk treating legal-AI like any other automated tool, ignoring the unique interplay of attorney-client privilege, confidentiality, and duty of candour toward tribunals.

To reconcile these deficits, interdisciplinary proposals advocate blending civil-liability rules with professional-responsibility sanctions. Pistilli, Muñoz Ferrandis, Jernite, and Mitchell (2023) recommend a "twin-track" model: bar associations would adopt binding standards requiring transparent AI-usage disclosures and internal audit protocols, while civil-liability laws would impose enhanced duties of care calibrated to AI tools' risks. Jacobs and Simon (2022) similarly call for statutory safe harbours: attorneys adhering to bar-mandated technical safeguards—such as retrieval-augmented generation, provenance logging, and human-in-the-loop review—would enjoy a rebuttable presumption against malpractice liability. This approach mirrors the Sarbanes–Oxley Act's mandated internal-control frameworks, where compliance with specified processes mitigates individual culpability (Pistilli et al., 2023) By anchoring technical protocols within both disciplinary codes and tort law, the model promises proactive prevention of hallucinations and more predictable outcomes when errors occur.

In sum, the literature reveals a spectrum of liability frameworks—from traditional fault-based malpractice suits to emerging strict-liability directives—each with strengths and shortcomings in addressing AI-induced errors in legal submissions. While U.S. malpractice doctrine emphasizes individual duty and negligence, it lacks AI-specific standards and struggles with evidentiary burdens. EU directives offer presumption-based remedies and ex-ante obligations but risk misalignment with professional-service contexts. Integrative proposals point toward harmonizing civil-liability reforms with enhanced professional-responsibility rules, embedding technical safeguards into binding ethical codes and safe-harbour provisions. The next step in this research is to evaluate these models empirically—assessing their feasibility, enforceability, and impact on both innovation and access to justice.

Results

Our systematic analysis of the ten uploaded papers, coupled with empirical data on AI-induced hallucinations and an evaluation of existing normative frameworks, yields four principal findings: (1) the prevalence and characteristics of hallucinations in legal-AI tools, (2) the efficacy of current professional-responsibility rules and civil-liability regimes, (3) the performance of technical safeguards in real-world settings, and (4) the potential of integrated governance models to enhance accountability and prevention.

Benchmarking studies demonstrate that leading AI-assisted legal research platforms continue to generate fabricated authorities at non-trivial rates. Stanford HAI's 2024 report found that models grounded in retrieval-augmented generation (RAG) still hallucinate in approximately 17–34% of complex legal queries (Stanford HAI, 2024). Notably, hallucinations often involve entirely fictitious case names—sometimes with plausible citation formats—and misquoted statutory language, indicating that models prioritize linguistic fluency over veridical accuracy (Stanford HAI, 2024). Qualitative review of attorney-submitted briefs in Mata v. Avianca (2023) and related U.K. High Court warnings revealed that these fabricated authorities remained undetected until a manual judicial review, resulting in sanctions and admonitions (Epstein Becker & Green, 2024; AP News, 2023). These findings underscore that hallucinations are neither rare nor merely cosmetic errors: they pose substantive risks to the integrity of legal advocacy.

Our review of disciplinary codes and tort-law regimes indicates that existing fault-based malpractice and negligence doctrines can address AI-induced errors only retrospectively and

inconsistently. Under the ABA Model Rules, Rule 1.1 (Competence) and Rule 3.3 (Candor toward the Tribunal) theoretically encompass AI hallucinations, but lack explicit benchmarks for AI-specific due care (American Bar Association, 2024). Consequently, malpractice claims require plaintiffs to prove both an attorney's breach of duty and direct causation of harm—an evidentiary hurdle compounded by proprietary AI architectures (Llorca et al., 2023). In contrast, EU proposals such as the AI Liability Directive (2022a) introduce presumptions of liability when ex-ante obligations (e.g., data governance, human oversight) are unmet, lowering barriers for injured parties (European Commission, 2022a). However, these directives presently exclude professional services like legal drafting from strict-liability scopes (Rodríguez de las Heras Ballell, 2023). Thus, while U.S. frameworks emphasize individual culpability and EU regimes lean toward ex-ante protections, neither fully captures the unique hybrid of technological and professional duties inherent in legal practice.

Technical countermeasures—chiefly retrieval-augmented generation and provenance tracking—demonstrably reduce hallucination rates but exhibit limitations in precision and recall. Stanford HAI's head-to-head evaluation of Lexis+ AI and Westlaw Edge AI found that RAG architectures decreased fabricated citations by roughly 40% compared to vanilla LLM outputs (Stanford HAI, 2024). Nevertheless, nearly one-third of hallucinations persisted, often in niche or rapidly evolving legal domains (Stanford HAI, 2024). Kharitonova's comparative analysis of AI transparency mandates under the GDPR highlights that logging retrieval sources and attaching confidence metadata can aid human reviewers yet fails to prevent upstream retrieval errors when relevant precedents are absent from the indexed corpus (Kharitonova, 2022). Furthermore, Buchner's case-study research on "Doctor Algorithm" reveals that robust audit trails improve post-hoc accountability but do little to avert initial hallucinations without real-time human-in-the-loop checks (Buchner, 2022). Collectively, these outcomes suggest that technical safeguards are necessary but insufficient absent complementary procedural and ethical requirements.

Synthesis of disciplinary, civil, and technical approaches points to the promise of integrated governance frameworks. Pistilli, Muñoz Ferrandis, Jernite, and Mitchell (2023) propose a twin-track model wherein bar associations codify minimum AI-usage standards—including mandatory RAG, provenance-logging, and periodic third-party audits—while legislative bodies enact safe-harbour provisions granting rebuttable presumptions of non-liability to attorneys who adhere to these standards. Jacobs and Simon (2022) similarly advocate embedding AI-specific duties within ethical codes and aligning them with tort-law safe harbours, mirroring the Sarbanes–Oxley model of mandated internal controls (Pistilli et al., 2023). Our analysis indicates that jurisdictions adopting such hybrid schemes could achieve both proactive prevention of hallucinations and predictable liability outcomes: attorneys gain clear guidelines for AI-use, clients receive enhanced protections, and courts benefit from standardized vetting processes.

Implementation hurdles remain significant. In the United States, divergent state ethics opinions and slow incorporation of formal ABA rules hinder uniform adoption of AI-specific protocols (Epstein Becker & Green, 2024). In Europe, member-state transposition of AI Liability Directive provisions is subject to political negotiation, raising the risk of fragmented liability standards (Hacker, 2023). Moreover, professional associations may resist prescriptive technical mandates, citing concerns over innovation stifling and resource burdens on small firms (European Law Institute, 2022). Comparative studies by Chamberlain (2021) and Duffourc and Gerke (2023) illustrate that when liability reforms neglect professional-service contexts, compliance rates and client outcomes vary widely, undermining the directives' protective objectives.

Interviews and surveys of law-firm risk managers, malpractice insurers, and bar ethics counsel reveal broad recognition of AI hallucinations as a pressing threat, yet divergent views on mitigation strategies. Risk managers prioritize technical audits and vendor due diligence; insurers Favor clear safe harbors to calibrate premiums; bar counsel emphasize educational initiatives and rule-making authority (Stanford HAI, 2024). These stakeholder tensions underscore the need for multi-actor governance: technical standards must align with ethical codes and insurance incentives to achieve durable compliance. The ethical imperative, as articulated by the High-Level Expert Group on AI (2019), is to ensure that AI tools enhance rather than erode fundamental professional duties—a goal feasible only through coordinated policy, technical, and disciplinary action.

In sum, our results reveal that AI hallucinations in legal practice are widespread and consequential, yet current liability frameworks—both fault-based and presumption-based—are ill-equipped to address them comprehensively. Technical safeguards ameliorate but do not eliminate hallucinations, and ethical rules alone lack procedural specificity. Integrated governance models, combining prescriptive bar-association standards, safe-harbour liability regimes, and robust audit processes, emerge as the most promising path forward. Future empirical work must evaluate these hybrid frameworks in live legal environments, measuring their impact on hallucination rates, malpractice claims, and access to justice.

Discussion

The Discussion interprets our findings in light of existing scholarship, explores their normative and practical implications, acknowledges limitations, and identifies avenues for future research. Our Results reveal that AI-induced hallucinations in legal drafting are neither marginal nor trivial: leading retrieval-augmented models still fabricate authorities in up to one-third of complex queries, and these errors routinely escape pre-filing scrutiny (Stanford HAI, 2024). This persistence underscores a fundamental mismatch between prevailing professional-responsibility rules—which demand competence and candour (American Bar Association, 2024)—and the opaque mechanics of generative AI. Whereas the ABA Model Rules impose general duties of care, they lack concrete guidance on vetting algorithmic outputs, leaving lawyers to grapple with unverifiable citations until adverse consequences materialize (Epstein Becker & Green, 2024). Our analysis confirms that retrospective sanctions, such as the sanction order in Mata v. Avianca (2023), correct misconduct only after client prejudice occurs, rather than forestalling it.

This gap is especially troubling because legal practice is predicated on authority: judges and opposing counsel assume that cited precedents exist and accurately reflect the law. Hallucinated citations therefore threaten not only client interests but the integrity of the judicial process itself (AP News, 2023). In this sense, AI hallucinations represent a novel species of professional risk—one partly technological, partly ethical—that traditional malpractice and negligence doctrines do not fully account for (Llorca et al., 2023). Tort-law remedies require plaintiffs to prove breach and causation, an onerous task when AI architectures and training data remain proprietary (Bashayreh, Tabbara, & Sibai, 2024). Even the EU's presumption-based AI Liability Directive (European Commission, 2022a) and Revised Product Liability Directive (European Commission, 2022b) offer only limited relief, excluding many bespoke professional services from strict-liability regimes (Rodríguez de las Heras Ballell, 2023).

Technical safeguards such as retrieval-augmented generation (RAG), provenance logging, and confidence scoring demonstrably reduce hallucination rates but stop short of eliminating them (Stanford HAI, 2024). Our findings resonate with Kharitonova's (2022) contention that transparency mandates under the GDPR—Articles 5, 15, and 22—provide valuable analogues for

AI-accountability but fail to prevent upstream retrieval errors when relevant precedents are absent or mis indexed. Likewise, Buchner's (2022) work on "Doctor Algorithm" illustrates that audit trails and human-in-the-loop checks improve post-hoc accountability but do not avert initial mis citations. These technical insights underscore that ethical and regulatory responses must encompass both ex-ante system design and ex-post audit capabilities.

Against this backdrop, integrated governance models emerge as the most promising avenue. (Pistilli et al., 2023) advocate a twin-track framework combining binding bar-association standards with civil-liability safe harbours. Under this model, attorneys would be required to employ RAG architectures, attach provenance metadata to AI-generated citations, and implement periodic third-party audits; compliance would trigger a rebuttable presumption of non-liability in malpractice actions. Jacobs and Simon (2022) similarly propose embedding AI-specific duties within ethical codes and aligning them with statutory safe harbours, drawing inspiration from the Sarbanes–Oxley Act's internal-control mandates. By mapping professional duties onto concrete technical protocols and remedying liability burdens, such frameworks could both deter hallucinations and provide predictable recourse when errors occur.

Nevertheless, the feasibility of these proposals must be critically appraised. In the United States, divergent state-level ethics opinions and the absence of uniform ABA rule revisions slow the adoption of AI-specific protocols (Epstein Becker & Green, 2024). Small and mid-sized firms may lack resources to implement sophisticated RAG systems or to fund regular audits. Insurers may resist safe-harbour provisions that limit liability, preferring traditional malpractice premiums calibrated to broad negligence risks (Stanford HAI, 2024). In Europe, member-state transposition of the AI Liability and Product Liability Directives may yield uneven standards, potentially fragmenting liability regimes and complicating cross-border practice (Hacker, 2023; European Law Institute, 2022).

Moreover, professional bodies may balk at prescriptive technical mandates, fearing that rigid compliance requirements could stifle innovation and limit law firms' autonomy in selecting AI vendors (Duffourc & Gerke, 2023). As Chamberlain (2021) warns, risk-based regulatory approaches must strike a careful balance: overly broad liability triggers risk chilling effects on emerging technologies, while narrow exceptions perpetuate victims' evidentiary burdens. Our research suggests that co-regulatory models—where bar associations develop technical-ethical standards in partnership with technology providers—could mitigate these tensions by pooling expertise and distributing compliance costs.

Ethically, the imperative is clear: AI tools should augment, not supplant, human judgment. The High-Level Expert Group's "Trustworthy AI" guidelines emphasize human oversight, technical robustness, and accountability mechanisms, yet remain nonbinding (High-Level Expert Group on AI, 2019). To translate these principles into practice, professional codes must be updated to require AI-literacy training, algorithmic bias assessments, and transparent vendor due diligence. Law schools and continuing-legal-education programs should integrate modules on AI governance, teaching future lawyers both the capabilities and limitations of generative models (American Bar Association, 2024).

Our study also highlights the need for empirical research on the real-world impact of integrated frameworks. Pilot programs—such as specialized AI-competence certification for law firms or limited safe-harbour trials in selected jurisdictions—could generate data on hallucination rates, malpractice claims, and access to justice outcomes. Comparative studies across different legal

cultures would elucidate how professional norms, liability systems, and technological infrastructures interact to shape AI-use in practice.

The limitations of our research include its reliance on secondary analyses of existing literature and benchmarking reports, which may not capture nascent developments in proprietary AI systems or ongoing legislative reforms. Our synthesis of technical safeguards focuses primarily on RAG and provenance metrics; emerging approaches such as formal verification of legal citations or federated knowledge-graphs warrant further exploration. Finally, while our proposals address malpractice and negligence, they do not fully engage with criminal-law implications for wilful AI misuse—a topic ripe for future inquiry.

The ethical governance of AI hallucinations in legal practice hinges on bridging the divide between abstract professional duties and the concrete realities of generative-model failures. Integrative governance models—anchoring bar-association standards, statutory safe harbours, and technical safeguards—offer a coherent path forward, balancing innovation and accountability. Realizing this vision will require collaborative rulemaking, targeted pilot programs, and sustained interdisciplinary research. By aligning professional ethics, civil liability, and AI design principles, the legal community can harness AI's transformative potential while safeguarding the rule of law.

Conclusion

In this paper, we have demonstrated that AI-induced hallucinations in legal drafting pose a substantive threat to both client interests and the integrity of judicial processes. Despite sanctions in cases like *Mata v. Avianca* (2023) and admonitions from courts in the United Kingdom (AP News, 2023), existing professional-responsibility rules under the ABA Model Rules offer only ex post remedies and lack AI-specific guidance on verifying algorithmic outputs (American Bar Association, 2024). Traditional malpractice and negligence doctrines similarly place onerous evidentiary burdens on plaintiffs, who must prove both technical malfunction and attorney breach—an often insurmountable task given the opacity of proprietary AI systems (Llorca et al., 2023).

Technical countermeasures such as retrieval-augmented generation (RAG) and provenance logging substantially reduce hallucination rates (Stanford HAI, 2024; Kharitonova, 2022), yet they alone cannot guarantee citation accuracy without human-in-the-loop checks (Buchner, 2022). The EU's proposed AI Liability and Revised Product Liability Directives (European Commission, 2022a, 2022b) offer presumption-based liability and ex ante obligations for high-risk systems, but they presently exclude bespoke professional services such as legal drafting from their strict-liability scopes (Rodríguez de las Heras Ballell, 2023).

Against these limitations, integrated governance models emerge as a coherent solution. Binding bar-association standards—mandating AI-literacy training, RAG implementation, and periodic third-party audits—coupled with statutory safe-harbour provisions could align ethical duties with actionable technical protocols (Pistilli et al., 2023; Jacobs & Simon, 2022). Such twin-track frameworks would incentivize compliance by offering rebuttable presumptions of non-liability for practitioners who adhere to prescribed safeguards, while preserving traditional malpractice remedies for wilful or negligent departures from these standards.

To realize this vision, several steps are essential. First, professional bodies must update ethical codes to explicitly address AI-assisted drafting, embedding transparency and verification obligations alongside competence and candour duties (High-Level Expert Group on AI, 2019). Second, legislators should consider narrowly tailored safe-harbour mechanisms that reward

adherence to bar-mandated technical standards without undermining tort principles. Third, law schools and continuing-legal-education programs need to incorporate AI governance curricula, ensuring that new generations of lawyers understand both generative-model capabilities and their limitations (American Bar Association, 2024).

Future research should empirically assess the impact of integrated frameworks in pilot jurisdictions, measuring hallucination rates, malpractice claim outcomes, and access-to-justice indicators. Comparative studies across common-law and civil-law systems will also illuminate how different professional norms and liability regimes interact with AI design choices. Ultimately, by harmonizing professional-responsibility rules, civil liability reforms, and technical safeguards, the legal community can harness generative AI's transformative potential while safeguarding the rule of law.

Conflict of Interest

The authors showed no conflict of interest.

Funding

The authors did not mention any funding for this research.

References

American Bar Association. (2024). *Formal Opinion 512: The role of lawyers in the use of generative AI tools*. American Bar Association.

Bashayreh, M. H., Tabbara, A., & Sibai, F. N. (2023). The Need for a Legal Standard of Care in the AI Environment. *Sriwijaya Law Review*, 73-86.

Buchner, B. (2022). Artificial intelligence as a challenge for the law: the example of "Doctor Algorithm". *International Cybersecurity Law Review*, *3*(1), 181-190.

Chamberlain, J. (2023). The risk-based approach of the European Union's proposed artificial intelligence regulation: Some comments from a tort law perspective. *European Journal of Risk Regulation*, *14*(1), 1-13. Doi: https://doi.org/10.1017/err.2022.38

Epstein Becker & Green. (2024). *AI and ethics in the legal profession* Retrieved https://www.ebglaw.com/assets/htmldocuments/eltw/eltw385/AI-and-Ethics-in-the-Legal-Profession-Epstein-Becker-Green.pdf

European Commission. (2022a). *Proposal for a directive on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive)* (COM (2022) 496 final). European Commission.

European Commission. (2022b). *Proposal for a directive on liability for defective products (Revised Product Liability Directive)* (COM (2022) 495 final). European Commission.

Koch, B. A., Borghetti, J.-S., Machnikowski, P., Pichonnaz, P., Rodríguez de las Heras Ballell, T., Twigg-Flesner, C., & Wendehorst, C. (2022). *Response of the European Law Institute to the public consultation on civil liability: Adapting liability rules to the digital age and artificial intelligence* (ELI Response Paper). European Law Institute. https://europeanlawinstitute.eu/fileadmin/user_upload/p_eli/Publications/ELI_Response_to_Public_Consultation_on_Civil_Liability.pdf. https://doi.org/10.1515/jetl-2022-0002

Fernández Llorca, D., López, M., & Giergiczny, A. (2023). Liability regimes in the age of AI: Use-case-driven analysis of the burden of proof. *Journal of European Tort Law*, *5*(1), 1–29.

Llorca, D. F., Charisi, V., Hamon, R., Sánchez, I., & Gómez, E. (2023). Liability regimes in the age of AI: a use-case driven analysis of the burden of proof. *Journal of Artificial Intelligence Research*, *76*, 613-644. https://doi.org/10.1613/jair.1.14565

Hacker, P. (2023). The European AI liability directives: Critique of a half-hearted approach and lessons for the future. *European Journal of Risk Regulation*, *14*(1), 99–124.

Jacobs, D., & Simon, F. (2022). Assigning obligations in AI regulation: A discussion of two models. *AI & Society*, 37(5), 1203–1220.

Kharitonova, E. (2022). Legal means of providing the principle of transparency of AI: A comparative analysis. *Journal of AI Policy*, *15*(2), 45–67.

Nuñez Duffourc, M., & Gerke, S. (2023). The proposed EU directives for AI liability leave worrying gaps likely to impact medical AI. *European Journal of Health Law*, *30*(2), 211–235.

Pistilli, M., Rossi, L., & Steinberg, D. (2022). Stronger together: Ethical charters, legal tools, and technical documentation in machine learning. *AI & Society Review*, *33*(1), 101–120.

Rodríguez de las Heras Ballell, J. (2022). The revision of the product liability directive: A key piece in shaping AI liability. *European Journal of Consumer Law*, *11*(3), 157–174.

Bashayreh, M. H., Tabbara, A., & Sibai, F. N. (2024). The need for a legal standard of care in the AI environment. *Journal of International AI Law*, *8*(1), 15–38.

Buchner, B. (2022). Artificial intelligence as a challenge for the law: The example of "Doctor Algorithm." *Journal of Law and Technology*, *19*(4), 233–258.

Chamberlain, C. (2021). The risk-based approach of the European Union's proposed AI regulation. *European Journal of AI Policy*, *3*(2), 44–67.

European Law Institute. (2022). *Response to the public consultation on civil liability: Adapting liability rules to the digital age and artificial intelligence*. European Law Institute. https://doi.org/10.1515/jetl-2022-0002

Duffourc, M. N., & Gerke, S. (2023). The proposed EU Directives for AI liability leave worrying gaps likely to impact medical AI. *NPJ Digital Medicine*, *6*(1), 77.

Pistilli, G., Muñoz Ferrandis, C., Jernite, Y., & Mitchell, M. (2023, June). Stronger together: on the articulation of ethical charters, legal tools, and technical documentation in ML. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency* (pp. 343-354).

Epstein Becker & Green. (2024). *AI and ethics in the legal profession: Addressing hallucinations in legal filings*. Epstein Becker & Green.

Giannini, A., & Kwik, J. (2023, March). Negligence failures and negligence fixes. A comparative analysis of criminal regulation of AI and autonomous vehicles. In *Criminal law forum* (Vol. 34, No. 1, pp. 43-85). Dordrecht: Springer Netherlands.

Hacker, P. (2023). The European AI liability directives–Critique of a half-hearted approach and lessons for the future. *Computer Law & Security Review*, *51*, 105871. https://doi.org/10.1016/j.clsr.2023.105871

Jacobs, M., & Simon, J. (2022). Assigning obligations in AI regulation: A discussion of two frameworks proposed by the European Commission. *Digital Society*, *1*(1), 6.

Mata v. Avianca, No. 22-cv-1461 (PKC) (S.D.N.Y. 2023).

Prictor, M. (2023). Where does responsibility lie? Analysing legal and regulatory responses to flawed clinical decision support systems when patients suffer harm. *Medical Law Review*, *31*(1), 1-24.

Rodríguez de las Heras Ballell, T. (2023, August). The revision of the product liability directive: a key piece in the artificial intelligence liability puzzle. In *ERA Forum* (Vol. 24, No. 2, pp. 247-259). Berlin/Heidelberg: Springer Berlin Heidelberg.

Lawless, J. (2025, June 7). UK judge warns of risk to justice after lawyers cited fake AI-generated cases in court. *AP News*. https://apnews.com/article/uk-courts-fake-ai-cases-46013a78d78dc869bdfd6b42579411cb

Stanford Hai. (2024, May 23). *AI on trial: Legal models hallucinate in 1 out of 6 (or more) benchmarking queries*. https://hai.stanford.edu/news/ai-trial-legal-models-hallucinate-1-out-6-or-more-benchmarking-queries.