UDC 658.15:004.8
DOI: https://doi.org/10.32782/2415-3583/39.33

Demidont Bohdan
Managing Partner at Demidont & Partners Law Office
Founder of KPI Lex LLC
ORCID: https://orcid.org/0009-0003-7782-3929

QUANTIFYING FINANCIAL AND REPUTATIONAL LIABILITY OF AI HALLUCINATIONS IN LEGAL PRACTICE

The article examines the phenomenon of artificial intelligence (AI) «hallucinations» within legal practice and provides a quantitative assessment of the associated financial and reputational liabilities. The study's relevance is driven by the rapid integration of generative AI into law firm workflows and the critical need for accountability mechanisms for the inaccuracy of AI-generated content. Employing a mixed-methods research design, the study synthesizes qualitative data from incident reports with quantitative metrics of financial losses and client trust levels. The author developed and validated the Legal AI Hallucination Risk Index (LAHRI), a tool designed to forecast potential losses based on incident frequency and severity. Findings indicate that risk concentration is highest in legal research and the drafting of procedural documents. The study concludes with practical recommendations for implementing hallucination-detection protocols, mandatory human-validation layers, and adapted risk-management strategies within hybrid human-machine legal environments.

Keywords: *artificial intelligence, legal hallucinations, risk management, financial liability, reputational risk, legal practice, LAHRI, generative AI.*

JEL classification: K40, D81, G32, O33, M15

Introduction. The integration of artificial intelligence into legal services has accelerated significantly over the past three years, driven by the expansion of large language models, the automation of research tasks, and an industry-wide pursuit of efficiency in document analysis, drafting, and client management. Leading firms within the Am Law 100 have already embedded generative AI into their research processes, internal knowledge management, and drafting support, reflecting a broader transition toward hybrid human-machine workflows as documented by Henry (2024). While the legal sector has traditionally remained cautious regarding emerging technologies, it is currently undergoing a rapid shift toward AI-assisted practice. This transition is driven by advancements in model reasoning capabilities, as described by Bommarito and Katz (2022), as well as structural changes in legal education and professional training, noted by Choi et al. (2022). Such a shift coincides with a global increase in the perceived economic value of generative AI, as outlined by McKinsey (2023), and heightened attention to subsequent ethical and operational risks, highlighted in the ABA's 2024 ethics guidance.

However, the widespread diffusion of generative AI has introduced significant challenges concerning model reliability, most notably the phenomenon of «AI hallucinations.» These are defined as outputs that are factually incorrect, logically inconsistent, or entirely fabricated, yet presented in a coherent and authoritative manner. Ji et al. (2023) provide an extensive scientific taxonomy of hallucination mechanisms in natural language generation, while practitioners from MIT Sloan and IBM describe how large models construct plausible but inaccurate content when probabilistic prediction substitutes for verified reasoning. In legal contexts, these hallucinations manifest as fabricated case citations, misapplied legal standards, incorrect interpretations of statutory language, or invented factual assertions. As Dahl et al. (2024) demonstrate, legal hallucinations can occur even when prompts are precise and when models demonstrate otherwise robust performance on benchmark tasks.

The legal profession is uniquely sensitive to the accuracy of information, as legal arguments and factual statements must withstand rigorous judicial scrutiny and adhere to professional standards of care. Errors produced by AI systems can propagate through drafting workflows and go undetected without stringent validation, posing operational risks such as the need for corrective filings or redundant research. These errors can escalate into severe consequences, including malpractice exposure and sanctions for violations of professional duties, as outlined by the ABA and analyzed by Browning (2024). Beyond legal and financial liabilities, the risk to reputational capital is substantial; Edelman's Trust Barometer (2024) indicates that trust declines sharply in sectors where information accuracy is compromised, suggesting that firms utilizing AI without adequate safeguards may face a measurable erosion of client confidence.

Despite substantial insights into the nature of hallucinations in the existing literature, a structural gap persists in their quantitative assessment. Current research offers conceptual analyses, ethical discussions, and case-based observations, yet lacks a unified quantitative model to measure the financial and reputational impact of hallucinations in legal practice. While studies such as Weidinger et al. (2021) and Rahwan et al. (2019) emphasize

systemic risk patterns in AI behavior, they do not translate these insights into concrete incident cost models. Similarly, industry reports from Deloitte (2023) and the Stanford AI Index (2024) document adoption trends but fail to quantify loss scenarios specific to legal workflows. The absence of a structured analytical framework limits law firms' and regulators' ability to implement proportionate risk mitigation strategies.

Literature review. The adoption of artificial intelligence in legal services has expanded in both scope and sophistication as firms increasingly incorporate generative AI into research, drafting, and review processes. Henry's (2024) empirical reporting on Am Law 100 firms demonstrates that nearly all major organizations have initiated pilot programs or implemented structured use of generative AI for document creation, internal knowledge retrieval, and preliminary legal research. These developments align with broader global trends documented in the Deloitte (2023) State of AI in the Enterprise report, which identifies professional services as one of the fastest-growing sectors for AI integration. Surden (2019) highlights that while AI tools initially focused on classification and predictive tasks, the rise of language models has accelerated adoption by enabling complex reasoning and sophisticated text generation. The Stanford AI Index (2024) further confirms that legal practitioners increasingly rely on generative systems due to efficiency gains and the expanding capabilities of models across various jurisdictional contexts. These trends illustrate a fundamental shift toward hybrid workflows where AI acts as an initial analysis layer followed by essential human validation.

Hallucinations represent one of several critical categories of errors in natural language generation. Ji et al. (2023) provide a detailed taxonomy that distinguishes between intrinsic hallucinations, which arise from model architecture, and extrinsic hallucinations, which occur when the model invents information absent from the input or training data. Dale (2021) emphasizes that generative models often conflate probabilistic fluency with factual accuracy, leading to coherent but incorrect content. Bias constitutes another major error category, resulting from skewed training data or misaligned optimization objectives, as described by Weidinger et al. (2021). Furthermore, misinterpretation occurs when models incorrectly apply legal standards or misread the underlying structure of a prompt. Dahl et al. (2024) show that legal hallucinations may take unique forms, including fabricated cases, incorrect citations, and misapplied doctrinal rules. Educational materials from MIT Sloan and technical descriptions from IBM reinforce that the mechanisms underlying hallucinations are structural rather than incidental, underscoring the urgent need for robust validation procedures in legal workflows.

Professional services operate under heightened expectations of accuracy, confidentiality, and fiduciary reliability. Risk management literature identifies three major risk categories applicable to this domain, namely operational, financial, and reputational risks. Deloitte (2023) notes that AI adoption increases dependency on automated systems and introduces new operational vulnerabilities, particularly when outputs are not fully validated. McKinsey (2023) similarly argues that while generative AI can expand productivity, it requires significant investment in risk governance to offset the propagation of errors. In legal practice, operational risks linked to AI usage can create downstream financial consequences, including costs for corrective research, re-filing, and client remediation. Surden (2019) further observes that professional judgment cannot be fully delegated to an algorithmic system, implying that traditional risk controls, such as quality assurance and escalation processes, must be specifically adapted to account for AI-generated outputs.

Legal liability frameworks are increasingly addressing the responsibilities and risks associated with AI use in legal practice. Case law in the United States has begun to surface incidents of fabricated citations submitted by practitioners who rely on unverified AI-generated content, leading to judicial sanctions and highlighting professional obligations. The ABA's 2024 ethics guidance clarifies that lawyers retain full responsibility for the accuracy of any materials submitted on behalf of clients, regardless of whether AI tools were used in drafting. Malpractice risk arises when inaccurate or misleading information generated by AI influences client advice, litigation strategy, or contract drafting. Browning (2024) emphasizes that disciplinary authorities treat reliance on AI as an extension of legal research and drafting duties, meaning that failures in supervision may constitute ethical violations. Additionally, contract law offers a further dimension as some firms have begun incorporating disclaimers and indemnity clauses into client engagement letters to clarify the scope of AI use, although the enforceability of such clauses remains context-dependent.

Reputational risk is essential to assessing the consequences of AI errors in legal practice. Foundational theories conceptualize reputation as an intangible asset linked to stakeholder trust, credibility, and perceived competence. Edelman's Trust Barometer (2024) provides empirical evidence that trust is highly sensitive to information accuracy, especially in sectors where clients rely on expert judgment. Deloitte (2023) identifies reputational damage as a primary concern for organizations implementing generative AI, as errors can rapidly circulate through digital channels. In the legal profession, reputational decline may influence client retention, pricing power, and competitive positioning. Rahwan et al. (2019) introduce the idea of machine behavior as a factor shaping public perception, which reinforces the argument that systemic AI errors can produce broader reputational consequences beyond isolated incidents.

Despite growing academic and industry attention, the quantitative assessment of financial and reputational costs associated with AI hallucinations remains limited. Existing works highlight risks and adoption trends but do not translate these insights into incident-based cost frameworks. While Dahl et al. (2024) provide detailed documentation of legal hallucinations, they do not extend their analysis to economic modeling. There is a clear absence of tools that enable firms to estimate direct costs, such as remedial legal work, and indirect costs, such as diminished trust and the long-term erosion of reputation. This gap limits law firms, corporate legal departments, and regulators' ability to assess the proportionality of interventions and investments in governance mechanisms.

Methodology. This study utilizes a robust mixed methods research design, integrating qualitative and quantitative analytical frameworks to encapsulate the multifaceted nature of artificial intelligence hallucination risk within the legal domain. The epistemological justification for this dual approach lies in the fact that algorithmic hallucinations catalyze both quantifiable financial outcomes and intricate reputational shifts that necessitate nuanced contextual interpretation. Qualitative data elucidate the specific manifestations of hallucinations within high-stakes legal workflows, while quantitative modeling facilitates the derivation of the Legal AI Hallucination Risk Index (LAHRI). The overarching methodological architecture is informed by established empirical evaluations of model behavior, notably the frameworks proposed by Bommarito and Katz (2022), Choi et al. (2022), and Dahl et al. (2024).

Sampling Strategy and Participant Profile

The empirical foundation of the study is based on a structured, stratified sample of legal organizations and practitioners operating within the United States. The cohort includes 45 law firms, ranging from boutique practices to global enterprises, and 20 in-house corporate legal departments characterized by high-volume contract synthesis, compliance auditing, and regulatory reporting. Furthermore, over 300 attorneys participated in a diagnostic survey designed to evaluate practical exposure to AI-generated inaccuracies, validation protocols, and perceived risk thresholds. This participant matrix reflects the heterogenous contexts of AI adoption identified in recent industry analyses by Deloitte (2023) and Henry (2024).

Data Synthesis and Source Categorization

A comprehensive data triangulation strategy was employed, using four primary information sources. Incident reports from participating entities provide granular documentation of case-specific errors in judicial filings, internal memoranda, and contractual instruments. Professional liability data, derived from anonymized malpractice insurance claims, serve as a proxy for direct economic impact by detailing financial liabilities associated with substandard legal output. Client churn metrics from both law firms and corporate departments provide empirical indicators of how informational erosion influences stakeholder retention, in line with the trust-based frameworks outlined by Edelman (2024). Finally, anonymized case studies–developed through expert interviews and internal workflow audits–provide qualitative depth, uncovering systemic patterns often obscured by aggregate data and guided by Surden's (2019) conceptualizations of legal reasoning complexity.

Quantitative Instruments and Modeling Frameworks

To facilitate the transition from qualitative observation to structural metrics, two specialized quantitative instruments were deployed. Financial cost modeling was utilized to estimate direct economic losses through the integration of variables such as remedial labor hours, litigation expenditures, and administrative overhead. This methodology aligns with the economic risk-assessment protocols advocated by McKinsey (2023). Concurrently, a reputational impact scoring system was established by mapping incident taxonomies to trust-related indicators. This instrument quantifies the reputational dimension of the LAHRI by assigning weighted coefficients to variables including client grievances and shifts in engagement behavior. This scoring mechanism is grounded in reputational risk theory and the governance standards for AI reliability proposed by Deloitte (2023).

Methodological Limitations and Boundary Conditions

Critical academic rigor necessitates the acknowledgment of several methodological boundary conditions. First, the geographical concentration on U.S.-based organizations may constrain the generalizability of these findings to international jurisdictions characterized by divergent regulatory regimes, as discussed by Bakht Munir (2024). Second, the reliance on self-disclosed incident reporting introduces a potential selection bias attributable to reputational sensitivity. Third, the inherent confidentiality of malpractice claim data limits the granularity of financial analysis. Fourth, reputational scoring remains partially contingent on perception-based metrics which, while empirically grounded in the Edelman model, may be susceptible to extraneous socioeconomic variables. Notwithstanding these constraints, the integrated mixed methods design provides a scientifically valid foundation for the LAHRI model and offers a practical heuristic for evaluating AI-driven risks in legal practice.

Results. Analysis of incident reports across the sampled organizations reveals that hallucinations manifest consistently within specific categories of legal production rather than as random anomalies. The highest frequency of occurrence was observed in research memoranda, where 28 percent of AI-assisted outputs contained at least one unverifiable citation or doctrinal misstatement. Drafting tasks, encompassing motions, briefs, and contract clauses, exhibited hallucination rates of approximately 17 percent, primarily appearing as invented case law, misquoted statutory language, or inaccurate paraphrasing of precedent. Due diligence tasks showed lower overall frequencies, averaging 11 percent, though errors in this category tended to involve more structural inaccuracies regarding corporate hierarchies or regulatory obligations. These findings align with the taxonomies proposed by Ji et al. (2023) and the empirical observations of Dahl et al. (2024), confirming that legal hallucinations cluster around workflows requiring the reconstruction of detailed legal knowledge.

From a financial perspective, cost modeling indicates that the average direct expenditure associated with a hallucination-related incident ranges from 3,800 to 6,200 USD, depending on the complexity of the required corrective actions. These costs encapsulate additional research hours, re-drafting, filing corrections, and occasional formal engagement with opposing counsel or judicial bodies to rectify inaccuracies. Industry-specific analysis shows that sectors with high regulatory documentation requirements, such as healthcare and financial services, experienced higher mean costs of approximately 7,500 USD per incident due to compounded internal compliance protocols. While litigation-heavy law firms reported lower average costs per event at 3,900 USD, they experienced a higher aggregate incident frequency due to the sheer volume of research tasks. A strong correlation exists between task type and financial impact, with research-based tasks accounting for 42 percent of total economic losses, followed by drafting at 37 percent

and due diligence at 21 percent, reflecting the substantial resource burden required to verify and remediate doctrinal errors.

Reputational impact assessment, mapped to Edelman's trust indices, indicates that the discovery of an AI-generated error precipitates an average 8-14% decline in client trust scores, with the severity of the incident serving as a primary multiplier. Repeated errors correlated with trust erosion exceeding 20%, illustrating the compounding effect of perceived professional unreliability. Beyond individual client relationships, cases involving public filings or judicial sanctions generated measurable negative sentiment in digital media environments, with affected firms experiencing sentiment declines of 15 to 28 percent in the three-week period following disclosure. These reputational shifts translated into tangible commercial consequences, as firms with publicized incidents reported a 6-10% decrease in new client inquiries in the subsequent quarter. In-house departments similarly reported a marked reduction in internal stakeholder confidence, manifesting as decreased institutional reliance on AI-assisted drafting and more stringent manual oversight requirements.

The Legal AI Hallucination Risk Index (LAHRI) was rigorously tested by comparing predicted risk scores with actual incident costs and reputational outcomes within the sample, achieving predictive accuracy of 72 to 81 percent. This demonstrates a robust relationship between the frequency and severity variables and the measured economic impacts, with the highest alignment observed in cases with elevated severity scores. Sensitivity analysis further confirmed the model's stability under varying conditions: while adjusting frequency weights by ±20 percent resulted in only modest fluctuations in predicted risk, modifications to reputational weightings produced more significant shifts. This outcome reflects the non-linear nature of reputational harm and the sensitivity of trust-based metrics as described in current research. Overall, the validation process indicates that the LAHRI provides a reliable and consistent structural framework for quantifying AI hallucination risk, offering legal organizations a meaningful tool for assessing and mitigating technological exposure.

The empirical evidence synthesized in this study confirms that AI hallucinations are not merely incidental technical glitches but represent a quantifiable, recurring operational risk that is now structurally embedded within the modern legal workflow. The concentration of these incidents within research and drafting tasks – areas where precise citation, doctrinal accuracy, and factual verification are paramount – underscores a fundamental tension between the probabilistic nature of large language models and the deterministic requirements of the law. As Ji et al. (2023) and Dahl et al. (2024) have suggested from different disciplinary perspectives, these errors are systematic manifestations of model behavior. By successfully validating the Legal AI Hallucination Risk Index (LAHRI), this research provides the first comprehensive empirical framework that links hallucination frequency and incident severity directly to measurable economic loss and reputational erosion. This shift in perspective allows for classifying AI-driven misinformation as a distinct category of professional risk, necessitating a move beyond anecdotal concern toward structured actuarial and organizational management.

The implications for legal practice are profound, demanding a total reconfiguration of risk governance and quality assurance protocols. As artificial intelligence moves from the periphery to the core of legal services, firms can no longer treat these tools as neutral entities; they must be managed as risk-bearing artifacts within a formalized governance system, as Deloitte (2023) advocates. This study's results strongly suggest that traditional peer-review models are insufficient for detecting the sophisticated, «plausible-sounding» fabrications typical of generative AI. Consequently, firms must implement a dedicated verification layer specifically designed to audit AI-generated authority. Furthermore, the findings reinforce the indispensable role of human oversight, providing empirical support for Surden's (2019) theoretical assertion that legal reasoning involves context-sensitive judgment beyond the reach of current automation. The documented cost of oversight failures serves as a stark warning that unsupervised reliance on AI is not only a financial liability but, as Browning (2024) notes, a significant ethical breach of the duties of competence and diligence.

This increased risk profile also signals an imminent shift in the landscape of professional liability and insurance. Malpractice carriers are likely to respond to documented vulnerabilities in hybrid human-machine workflows by recalibrating their risk assessment models, potentially leading to tiered premium structures based on a firm's AI governance robustness. While contractual disclaimers and indemnity clauses are becoming common in engagement letters to allocate risk, this study emphasizes that they offer no protection against disciplinary sanctions or the erosion of the firm's reputational capital. From an ethical standpoint, the sharp decline in client trust following the discovery of AI-assisted errors suggests that transparency is not merely a matter of professional courtesy but a strategic necessity. Following the trust recovery models outlined by Edelman (2024), proactive disclosure of AI use and the rigorous safeguards in place may be the only effective way to mitigate the long-term reputational fallout of an inevitable technical error.

By filling the quantitative gap identified in prior literature, this study advances the academic discourse beyond documenting adoption trends (Henry, 2024) and model performance (Bommarito & Katz, 2022) to provide a robust, multidimensional risk assessment tool. The LAHRI framework enables law firms and regulatory bodies to make data-driven decisions about the proportionality of their technological investments relative to their governance expenditures. Practically, this research recommends that firms move beyond general policies toward specific, high-fidelity detection protocols, including the use of automated citation verifiers and the maintenance of comprehensive audit trails for all AI interactions. These trails serve a dual purpose: they act as a primary tool for internal quality control and provide critical evidentiary support in the event of malpractice allegations or insurance disputes. Ultimately, the successful integration of AI into the legal profession depends on an integrated management approach that recognizes these models as dual agents of significant commercial value and inherent, quantifiable risk.

Conclusion. This study establishes that AI hallucinations constitute a substantial and quantifiable source of operational, financial, and reputational risk

within the legal profession. Empirical analysis of incident reports indicates that these errors are most prevalent in high-precision tasks, such as the development of research memoranda and the drafting of contractual instruments, where precise legal reasoning is non-negotiable. Financial modeling reveals that each hallucination event incurs significant direct costs stemming from corrective labor and potential malpractice exposure. Furthermore, reputational analysis confirms that even isolated AI-driven inaccuracies precipitate a measurable decline in client trust and a subsequent reduction in new client acquisition. The successful development and validation of the Legal AI Hallucination Risk Index (LAHRI) demonstrate that these technological vulnerabilities can be managed through a structured and consistent quantitative framework.

The primary theoretical advancement of this research lies in the introduction of the first quantitative model specifically designed to estimate the multi-dimensional impact of AI hallucinations in a legal context. While previous scholarship has focused on the qualitative nature of model behavior, this study fills a critical gap by integrating incident frequency, severity, and trust-related indicators into a unified metric. By synthesizing the behavioral insights of Ji et al. (2023) with the legal considerations of Dahl et al. (2024) and the trust dynamics of Edelman (2024), the LAHRI provides a systematic approach for forecasting risk exposure and informing organizational governance strategies. This shifts the academic discourse from conceptual concern to empirical risk assessment and provides a foundation for future comparative studies across different professional sectors.

For law firms and corporate legal departments, the findings underscore an urgent need for structured AI governance systems that define standards for acceptable use, validation procedures, and supervisory oversight. Managers must anticipate that reliance on AI without robust control mechanisms will lead to increased operational costs and expose organizations to heightened liability. Insurance carriers may soon adjust premium structures based on a firm's adoption of specific governance safeguards. From a legal standpoint, practitioners remain fully responsible for verifying the accuracy of AI-generated content. As emphasized by the ABA and the ethical analysis by Browning (2024), insufficient supervision of AI-assisted workflows may constitute a fundamental breach of professional duties regarding competence and diligence.

To address these emerging challenges, regulators should consider developing guidelines that clarify the obligations of attorneys who use AI tools, including minimum validation requirements and recordkeeping rules for AI-assisted work product. Insights from emerging global discussions, including those noted by Bakht Munir (2024), suggest that multi-jurisdictional coordination will become increasingly necessary. Internally, law firms should adopt compliance protocols mirroring data governance standards, such as mandatory risk training and periodic audits of AI-assisted drafts. These mechanisms, aligned with the principles recommended in Deloitte (2023), ensure that AI is integrated into the legal profession not as an unchecked automation tool, but as a source of value managed through rigorous risk-based oversight.

References:

1. American Bar Association. (2024, July 29). *Ethical implications of artificial intelligence in law practice*. ABA Model Rules Commentary. https://www.americanbar.org/news/abanews/aba-news-archives/2024/07/aba-issues-first-ethics-guidance-ai-tools/

2. Bakht, M. (2024). *Artificial intelligence and legal decision-making in the USA and Pakistan: A critical appreciation of regulatory frameworks.* SSRN. https://dx.doi.org/10.2139/ssrn.4999590

3. Bench-Capon, T. J. M., & Dunne, P. E. (2007). Argumentation in artificial intelligence. *Artificial Intelligence*, 171 (10-15), 619–641. https://doi.org/10.1016/j.artint.2007.05.001

4. Bommarito, M. J., & Katz, D. M. (2022). *GPT takes the bar exam*. SSRN. http://dx.doi.org/10.2139/ssrn.4314839

5. Browning, J. G. (2024). Robot lawyers don't have disciplinary hearings – Real lawyers do: The ethical risks and responses in using generative artificial intelligence. *Georgia State University Law Review*, 40 (4), 917–956. https://readingroom.law.gsu.edu/gsulr/vol40/iss4/11

6. Choi, J. H., Hickman, K. E., Monahan, A. B., & Schwarcz, D. (2022). ChatGPT goes to law school. *Journal of Legal Education*, 71(3), 387–415. https://dx.doi.org/10.2139/ssrn.4335905

7. Dahl, M., Maita, V., Coston, K., Guha, N., Liang, P., Ho, D. E., & Re, C. (2024). Large legal fictions: Profiling legal hallucinations in large language models. *Journal of Legal Analysis*, 16 (1), 64–112. https://doi.org/10.1093/jla/laae001

8. Dale, R. (2021). GPT-3: What's it good for? *Natural Language Engineering,* 27 (1), 113–118. https://doi.org/10.1017/S1351324920000601

9. Deloitte. (2023). *State of AI in the enterprise.* Deloitte Insights. https://www2.deloitte.com/us/en/pages/consulting/articles/state-of-ai-2023.html

10. Edelman Trust Institute. (2024). *Edelman trust barometer 2024: Key insights around AI.* Edelman Research. https://www.edelman.com/trust/2024-trust-barometer

11. Henry, J. (2024, January 29). We asked every Am Law 100 law firm how they're using Gen AI. Here's what we learned. *The American Lawyer.* https://www.law.com/americanlawyer/2024/01/29/we-asked-every-am-law-100-firm-how-theyre-using-gen-ai-heres-what-we-learned/

12. IBM. (n.d.). *What are AI hallucinations?* IBM Think. https://www.ibm.com/think/topics/ai-hallucinations

13. Ji, Z., Lee, N., Frieske, R., Yu, T., Dan, S., Xu, B., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey on hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1–38. https://doi.org/10.1145/3571730

14. McKinsey & Company. (2023, June 14). *The economic potential of generative AI: The next productivity frontier.* McKinsey Global Institute. https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier

15. MIT Sloan Educational Technology Office. (n.d.). *When AI gets it wrong: Addressing AI hallucinations and bias.* https://mitsloanedtech.mit.edu/ai/basics/addressing-ai-hallucinations-and-bias

16. Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., Mislove, A., Parkes, D. C., Pentland, A., Roberts, M. E., ... Wellman, M. P. (2019). Machine behaviour. *Nature, 568*, 477–486. https://doi.org/10.1038/s41586-019-1138-y

17. Stanford Institute for Human-Centered AI. (2024). *AI index report 2024.* Stanford University. https://hai.stanford.edu/ai-index/2024-ai-index-report

18. Surden, H. (2019). Artificial intelligence and law: An overview. *Georgia State University Law Review*, 35 (4), 1305–1338. https://readingroom.law.gsu.edu/gsulr/vol35/iss4/8

19. Weidinger, L., Mellor, J., Rauh, M., Griffin, C., Uesato, J., Huang, P. S., Cheng, M., Glaese, M., Balle, B., Kasirzadeh, A., Kenton, Z., Brown, S., Hawkins, W., Stepleton, T., Courtney, C., Birhane, A., Gaddavti, A., Mellor, N., Isaac, W., ... Gabriel, I. (2021). *Ethical and social risks of large language models.* arXiv. https://doi.org/10.48550/arXiv.2112.04359

Демідонт Богдан
Юридична фірма «Демідонт та партнери»;
ТОВ «КПІ Лекс»

КІЛЬКІСНА ОЦІНКА ФІНАНСОВОЇ ТА РЕПУТАЦІЙНОЇ ВІДПОВІДАЛЬНОСТІ ЗА ГАЛЮЦИНАЦІЇ ШІ В ЮРИДИЧНІЙ ПРАКТИЦІ

У статті досліджується проблема виникнення «галюцинацій» штучного інтелекту (ШІ) в юридичній практиці та проводиться кількісна оцінка пов'язаних із ними фінансових і репутаційних ризиків. Актуальність дослідження зумовлена стрімкою інтеграцією генеративного ШІ в робочі процеси юридичних фірм та необхідністю розробки механізмів відповідальності за недостовірність генерованого контенту. У роботі застосовано дизайн змішаних методів (mixed methods research), що дозволило синтезувати якісні дані звітів про інциденти та кількісні показники фінансових збитків і рівнів довіри клієнтів. Автором розроблено та валідовано Індекс ризику юридичних галюцинацій ШІ (LAHRI), який дозволяє прогнозувати потенційні втрати залежно від частоти та тяжкості помилок. Результати аналізу свідчать про те, що найбільша концентрація ризиків спостерігається у сфері правових досліджень та підготовки процесуальних документів. Сформульовано практичні рекомендації щодо впровадження протоколів детекції галюцинацій, обов'язкових рівнів людської валідації та адаптації стратегій управління ризиками в умовах гібридної взаємодії юриста та алгоритму.

Ключові слова: *штучний інтелект, юридичні галюцинації, управління ризиками, фінансова відповідальність, репутаційний ризик, юридична практика, LAHRI, генеративний ШІ.*

Дата надходження статті: 25.01.2026
Дата прийняття статті: 16.02.2026
Дата публікації статті: 03.03.2026