10. Angwin J., Larson J., Mattu S., Kirchner L. Machine Bias. *ProPublica* . 2016. 23 May. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (дата звернення: 08.06.2026).

11. Ієнка М., Андорно Р. На шляху до нових прав людини в епоху нейротехнологій: когнітивна свобода, ментальна конфіденційність, ментальна цілісність та психологічна безперервність. *Науки про життя, суспільство та політика* . 2017. Том 13, № 5. DOI: https://doi.org/10.1186/s40504-017-0050-1 .

12. Haber E. The Criminal Metaverse. *Indiana Law Journal*. 2024. Vol. 99, iss. 3. P. 843–891. URL: indiana.edu (дата звернення: 08.06.2026).

13. State v. Loomis. 371 Wis. 2d 235, 881 N.W.2d 749 (Wis. 2016). URL: https://www.wicourts.gov/sc/opinion/DisplayDocument.pdf?content=pdf&seqNo=171690 (дата звернення: 08.06.2026).

Tim D. WILLIAMS
BSc, MSc, MBA, PGCHE
FBCS CITP, MCIIS, MIET, GMBPsS
Independent researcher
ORCID: 0000-0002-1788-5736

WHY AI LEGAL ASSISTANTS HALLUCINATE IN NEW FIELDS: A MATHEMATICAL LIMIT, WITH POST-QUANTUM CRYPTOGRAPHY (PQC) AS EXAMPLE

1. Introduction and motivation

Some AI errors in legal practice cannot be removed by better engineering. They follow from a mathematical limit. Practitioners need to know which legal questions sit inside that limit.

AI assistants are now part of everyday legal work. Lawyers use them for research, drafting and first-pass review. When these assistants make confident mistakes, as in the US case Mata v. Avianca [1], the usual response is to call for better training data, better prompts and better engineering. This paper offers a different perspective. One specific kind of AI error cannot be removed by engineering. It follows from a mathematical proof. The proof applies to any AI assistant that gives confidence-marked answers, both now and in the future. The professional task is therefore not to wait for AI to become reliable. The task is to learn which legal questions fall inside the mathematical limit, and to avoid reliance on AI outside its known limits of reliability. Post-quantum cryptography (PQC) is used here as a worked example. PQC is a new legal field. The public record is thin. The stakes are high.

2. The Kalai–Vempala lower bound

A 2024 mathematical result shows that any properly calibrated AI assistant must make a minimum rate of errors. The minimum rate depends on how often facts appear in the training data.

In 2024, Kalai and Vempala proved a result about AI language models [2]. The result is a lower bound, which means a minimum that cannot be reduced. This bound applies to any AI model that is well-calibrated, meaning a model whose stated confidence matches how often it is correct. This bound says: the model must make errors at a rate at least equal to the share of facts that appear exactly once in its training data. Xu, Jain and Kankanhalli then strengthened the result [3]. They showed that some errors are unavoidable as a matter of computability, not training effort.

The practical meaning is simple. An AI assistant is not equally reliable across all domains of knowledge. The model is reliable where the same fact appears many times, in many sources. The model is unreliable where the authoritative source appears only once. This is not a problem of current models. It is a proven limit on every future calibrated model. The professional task is to identify which legal domains fall on the unreliable side of the limit.

3. A second factor: gaps in the legal training record

The mathematical limit is an underestimate where AI assistant training data is not only limited but also distorted. Legal practice has structural reasons for such distortion.

The Kalai–Vempala limit assumes that thin training data is the only problem. In legal practice, the data is also distorted. Anderson showed that banks for many years suppressed reports of cryptographic failure [4]. The public record was therefore biased towards success and away from instructive failure. Legal practice has the same structural feature. Sealed settlements, confidential dispositions, protective orders and legal professional privilege all remove the most instructive failure cases from the public record [5]. These are exactly the cases an AI assistant would need in order to learn. The combined effect is that AI legal assistants are under-exposed to the newest and most relevant subject matter and over-exposed to favourable outcomes. Anderson's earlier economic account of cryptographic failure [6] and this author's earlier work on epistemic grounding in cybersecurity [7] both support the same conclusion. Confident error is the predictable result. The mechanism and the economic conditions in which it operates are analysed by this author in a forthcoming paper [8].

4. Why PQC-legal advisory is the canonical case

PQC-legal advisory is the clearest near-term field where the Kalai–Vempala limit applies most acutely. PQC combines all three features that trigger the limit.

Post-quantum cryptography has moved from a research topic to a live legal field. In August 2024 the US National Institute of Standards and Technology approved the first three PQC standards [9]. This opened a regulatory and advisory market that cuts across financial reporting [10], operational resilience [11], and emerging negligence claims [12]. Lawyers in many roles, including in-house counsel, advocates, notaries and judges, now use AI assistants to read this material at speed. The problem is that PQC-legal advisory sits exactly where the Kalai–Vempala limit predicts trouble. No organisation is known to have finished a full PQC transition. No comprehensive public cost data exists [13]. The engineering solutions continue to change. Three features converge: (1) the subject is recent; (2) the boundary between technical and legal expertise is unsettled; and (3) the legal questions are jurisdiction-specific. Each feature is hard on its own. The three together compound. PQC is perhaps the first such field on which the legal profession will be tested. Other novel fields with similar legal/AI features include AI liability, quantum sensing, synthetic biology and autonomous-systems law.

5. A proposed empirical test

This paper's theoretical argument supports a small empirical test that any research team can run as follows.

A set of 30 to 40 prompts on PQC-legal questions is drawn from NIST guidance, IAS 36 application material, DORA and NIS2 provisions, and recent sanctions decisions. A matched set of prompts of similar difficulty is drawn from settled IT-legal subject matter. Both sets are run against three frontier AI models under the same conditions. Two raters then grade the responses against authoritative sources, blinded to condition where this is possible. The primary expectation is that the PQC-legal prompts produce a much higher rate of confident error.

Five categories of error are scored: fabricated authority, misattributed holding, incorrect standard reference, confidently wrong timeline, and unsupported doctrinal inference. The test is diagnostic, not decisive. It cannot disprove the Kalai–Vempala result. It can only show the predicted pattern in the PQC-legal condition. The paper's main contribution is the structural explanation. The measured size of the effect is secondary. This pilot is intended to support a later journal extension.

6. What this means for legal practice

If AI error in legal advisory is mathematically structured rather than uniform, the professional response should change in specific ways.

The duty on counsel is not to eliminate a general defect. The duty is to identify the high-risk scenarios where legal AI assistants will confidently generate incorrect results. From the earlier sections, three signals of high-risk scenarios support a practical framework: recent subject emergence,

jurisdiction-specific novelty, and technical–legal cross-domain character. Where two or more signals are present, counsel should presume an elevated risk of AI error. Disclosure and verification duties should follow that presumption in proportion.

Three practical implications follow. For in-house counsel and notaries, structured verification protocols should apply to any AI output in the flagged scenarios, with specific checks for fabricated standards and misattributed cases. For judges, protocols for AI-assisted filings should treat the medium as neither uniformly reliable nor uniformly suspect. For bar associations and regulators, domain-specific guidance can be issued now, without waiting for a general solution; the structural account supplies the criteria for naming the domains. A final note is required on retrieval-augmented generation (RAG) [14]. RAG shifts the sparsity problem rather than solving it. Where the retrievable authoritative material is itself sparse, RAG offers no relief [15].

7. Conclusion

AI error in legal practice is partly structural and partly an engineering problem. The structural part can be named and managed. AI error in legal assistants is not uniform across legal fields. One specific class of error is mathematically structural, and the distortion of the legal training corpus makes it worse. PQC-legal advisory is the clearest near-term example. The diagnosis generalises to every emerging-technology field in which counsel is asked for early advice. The professional task is not to wait for AI reliability to improve. The task is to recognise the regime in which reliability will remain bounded by mathematics, and to respond with proportionate verification and disclosure.

Conflict of interest disclosure

The author is engaged in commercial PQC advisory activity within the financial-services sector, and is co-founder of ProteQC, a firm in which Darren Bender (cited at reference 12) serves as Chief Legal Officer. No commercial material informs this submission. All argumentation is based on publicly available peer-reviewed sources, public regulatory texts and public court records.

References (DSTU 8302:2015)

1. Mata v. Avianca, Inc., No. 22-cv-1461 (PKC), United States District Court, Southern District of New York. Order imposing sanctions, 22 June 2023. URL: https://storage.courtlistener.com/recap/gov.uscourts.nysd.575368/gov.uscourts.nysd.575368.54.0_7.pdf (accessed: 22.04.2026).

2. Kalai A. T., Vempala S. S. Calibrated language models must hallucinate // Proceedings of the 56th Annual ACM Symposium on Theory of Computing (STOC 2024). New York : ACM, 2024. P. 160–171.

3. Xu Z., Jain S., Kankanhalli M. Hallucination is inevitable: an innate limitation of large language models. 2024. arXiv:2401.11817. URL: https://arxiv.org/abs/2401.11817 (accessed: 22.04.2026).

4. Anderson R. Why cryptosystems fail // Proceedings of the 1st ACM Conference on Computer and Communications Security. New York : ACM, 1993. P. 215–227.

5. Galanter M. Why the “haves” come out ahead: speculations on the limits of legal change // Law & Society Review. 1974. Vol. 9, No. 1. P. 95–160. DOI: 10.2307/3053023.

6. Anderson R. Why information security is hard — an economic perspective // 17th Annual Computer Security Applications Conference (ACSAC 2001). New Orleans : IEEE, 2001. P. 358–365.

7. Williams T. D. Epistemological questions for cybersecurity // 2020 International Conference on Cyber Security and Protection of Digital Services (Cyber Security). Dublin : IEEE, 2020. P. 1–4. DOI: 10.1109/CyberSecurity49315.2020.9138884.

8. Williams T. D. Cybersecurity economics, PQC and AI — 25 years post Anderson (2001) // Proceedings of Cyber Science 2026. London : Springer LNCS, 2026. [in press].

9. National Institute of Standards and Technology. NIST releases first 3 finalized post-quantum encryption standards. Gaithersburg : NIST, 13 August 2024. URL: https://www.nist.gov/news-events/news/2024/08/nist-releases-first-3-finalized-post-quantum-encryption-standards (accessed: 22.04.2026).

10. IAS 36 Impairment of Assets. London : International Financial Reporting Standards Foundation, as amended. URL: https://www.ifrs.org/issued-standards/list-of-standards/ias-36-impairment-of-assets/ (accessed: 22.04.2026).

11. Regulation (EU) 2022/2554 of the European Parliament and of the Council of 14 December 2022 on digital operational resilience for the financial sector (DORA) // Official Journal of the European Union. 2022. L 333/1–L 333/79.

12. Bender D. Post-Quantum Negligence: applying the Learned Hand formula to organizational liability for quantum computing threats. SSRN Working Paper No. 6413418, March 2026. DOI: 10.2139/ssrn.6413418. URL: https://ssrn.com/abstract=6413418 (accessed: 21.05.2026).

13. Williams T. D. Towards a framework for post-quantum cryptography cost estimation. SSRN Working Paper No. 6726106, May 2026. DOI: 10.2139/ssrn.6726106. URL: https://ssrn.com/abstract=6726106 (accessed: 21.05.2026).

14. Lewis P., Perez E., Piktus A. et al. Retrieval-augmented generation for knowledge-intensive NLP tasks // Advances in Neural Information Processing Systems 33 (NeurIPS 2020). Red Hook, NY : Curran Associates, 2020. P. 9459–9474. arXiv:2005.11401.

15. Kandpal N., Deng H., Roberts A., Wallace E., Raffel C. Large language models struggle to learn long-tail knowledge // Proceedings of the 40th International Conference on Machine Learning (ICML 2023). PMLR, 2023. Vol. 202. P. 15696–15707.

Ірина ГЛОВЮК, професорка кафедри кримінального процесу та криміналістики, Одеський державний університет внутрішніх справ, доктор юридичних наук, професор

ШІ ТА ЗАКОН УКРАЇНИ «ПРО АКАДЕМІЧНУ ДОБРОЧЕСНІСТЬ»: ПОДАЛЬШІ ПИТАННЯ

Одним з новітніх викликів у аспекті академічної доброчесності є використання ШІ у науковій та освітній діяльності. За ст. 29 Закону України «Про академічну доброчесність», недоброчесне використання результатів, згенерованих штучним інтелектом, - це оприлюднення текстів, зображень, моделей, даних, інших результатів, згенерованих штучним інтелектом, як результатів власної академічної діяльності, якщо цей факт не зазначено в академічному творі чи супровідних матеріалах до нього. У ч. 4 ст. 8 цього ж Закону прописано, що у разі використання в академічному творі об'єкта, згенерованого штучним інтелектом, автор повинен повідомити про це в такому творі із зазначенням методики генерування та/або посилання на відповідну комп'ютерну програму чи її опис згідно з вимогами щодо оформлення та/або оприлюднення відповідних академічних творів, визначених відповідним суб'єктом академічної діяльності.

Ці норми є більш прогресивними, аніж ті, що попередньо пропонувалися у проєкті цього Закону. Адже у ст. 8 проєкту вказувалося, що особа не може вважатися автором академічного твору (частини академічного твору), якщо він сформований (згенерований) за запитом особи комп'ютерною програмою в автоматичному режимі. Додаткові питання виникають, наприклад, якщо ШІ допоміг оформити перелік джерел (наприклад, переформатував у потрібний стиль цитування): частина тексту згенерована ШІ, отже, якщо формально підходити до цього питання, то це порушення академічної доброчесності. А якщо ШІ лише сформулював тему? А якщо лише план? А якщо вичитав і вказав, як краще переформулювати 2-3 вже написані людиною фрази? Відповідей немає. Утім, є положення: «При використанні в академічному творі частин, сформованих (згенерованих) комп'ютерними програмами, цей