Artificial Intelligence, Hallucinated Citations, and the Risk of Self-Validating Knowledge

Helena Donato[1,2*]

1. Serviço Documentação e Informação Científica, Hospitais da Universidade de Coimbra, Unidade Local de Saúde de Coimbra, Coimbra, Portugal.
2. Portuguese Kidney Journal (PKJ) Editing Consultant

https://doi.org/10.71749/pkj.155

INTRODUCTION

Artificial Intelligence (AI) is frequently portrayed as a sudden, existential threat to the integrity of scientific publishing. However, the reality is more nuanced: AI did not create the existing fragilities within the scientific system but has instead served as a powerful catalyst, exacerbating pre-existing issues.[1] Practices such as the fraudulent production of articles, data fabrication, plagiarism, citation manipulation, the purchase of authorship, and the systemic pressure to "publish or perish" were well-established long before the widespread adoption of AI tools.[2]

The primary change introduced by AI is the democratization and acceleration of these harmful practices. AI has made the production of deceptive content faster, less expensive, and significantly more difficult for editors and reviewers to control.[2] As large language models (LLMs) become integrated into the research workflow, the scientific community faces a fundamental shift in how knowledge is produced and validated.

AI-GENERATED MANUSCRIPTS AND THE RISE OF HALLUCINATION

The current generation of LLMs possesses the capability to generate comprehensive literature reviews, draft entire manuscripts, and produce convincing bibliographies within seconds.[1,2] While these capabilities offer efficiency, they also introduce a critical flaw: "hallucination." Many of the references generated by these models are entirely false, while others may exist in reality but do not correspond to the content they are cited to support.[2] Scientific integrity has historically relied on a simple, foundational assumption: that which is cited exists, has been read by the authors, and effectively supports the claim being made.[2] In the era of AI-generated content, this assumption can no longer be taken for granted. The presence of fabricated references is not merely a technical error but a direct challenge to the veracity of the scientific record.

THE MECHANICS OF DECEPTION: FRANKENSTEIN CITATIONS

The deceptive nature of AI-generated references lies in their plausibility. These fabrications often appear credible because they are constructed using authentic elements harvested from real academic papers. Research published in Nature describes these as "Frankenstein citations": hybrid constructions where real authors, plausible titles, existing journals, and partially correct digital object identifiers (DOIs) are blended to create a fictional but convincing citation.[3,4] Because these citations use familiar names and legitimate journal titles, they can easily bypass a casual inspection. They are correctly formatted and attributed to active researchers, making the task of identifying them as fabrications labor-intensive for human editors and reviewers.[3]

EMPIRICAL EVIDENCE FROM MAJOR JOURNALS

The scale of the hallucinated citation problem has been documented by recent large-scale audits. A significant study published in The Lancet analyzed 2.5 million biomedical articles to identify the prevalence of fabricated references.[3] This audit revealed thousands of fabricated citations, many of which were formatted with such precision and attributed to such plausible researchers that they were indistinguishable from real references without deep verification.[4] Similarly, an analysis titled "The Growing Problem of Hallucinated Citations" in Nature highlights how bibliographies that are either entirely invented or significantly adulterated by AI models have become a concrete threat to scientific integrity.[4] The rapid growth

Received: 23/06/2026 **Accepted:** 24/06/2026 **Published Online:** 25/06/2026 **Published:** 25/06/2026

* **Corresponding Author:** Helena Donato | helenadonato@ulscoimbra.min-saude.pt | Directora Serviço Documentação e Informação Científica, Hospitais da Universidade de Coimbra, Unidade Local de Saúde de Coimbra, Av. Bissaya Barreto, 3000-075 Coimbra

© PKJ 2026. Re-use permitted under CC BY-NC 4.0. (https://creativecommons.org/licenses/by/4.0/)

of this phenomenon suggests that the pollution of scientific literature with non-existent data and references is an escalating crisis.[4]

ETHICAL CONSEQUENCES AND REPUTATIONAL RISKS

The inclusion of "hallucinated" references is classified as a serious violation of scientific integrity. The consequences of such violations depend on when they are discovered:

- Editorial Phase: If fabricated references are detected during the editorial or peer review process, they are typically viewed as a potential indicator of broader scientific fraud. In such cases, the manuscript is generally rejected outright.[2]
- Post-Publication: If these errors are only identified after the work has been published, they often lead to the formal retraction of the article.[2]

Retractions carry a long-lasting reputational impact for the authors involved, compromising the perceived credibility of their entire body of work and potentially harming their professional standing within the scientific community.[2]

THE CLOSED-LOOP CRISIS: A SELF-REFERENTIAL ECOSYSTEM

The most significant risk posed by AI in publishing is the potential for a "closed loop" in knowledge production. We are facing a scenario where:

1. AI writes the articles.
2. AI assists in or conducts the review of those articles.
3. AI summarizes the resulting published content.
4. AI models are subsequently trained on this unvalidated, machine-generated content.[2]

This "closed loop" threatens to create a self-validating system of knowledge that may drift further and further away from empirical truth.[2] Although this scenario may seem extreme, the fundamental elements for its realization are already present in the current publishing landscape.

The risks associated with AI in scientific publishing go beyond fabricated references and inaccurate citations. A less visible but potentially more serious challenge is the growing ability of AI systems to generate linguistically sophisticated, structurally coherent, and seemingly credible manuscripts that are, however, methodologically weak.[6] By drastically reducing the time and specialized knowledge required to produce scientific texts, AI lowers the barriers to manuscript production and may inadvertently amplify existing pressures within the academic system, particularly the incentives created by the "publish or perish" culture.[7]

This development raises concerns about the growing volume of studies that appear scientifically credible but are based on flawed methodologies, inadequate study designs, analyses with insufficient statistical power, or clinically irrelevant results. Consequently, the scientific literature risks becoming increasingly filled with articles that are technically publishable but contribute little meaningful knowledge.

Without rigorous methodological oversight and critical interpretation, this capability may prioritize quantity over quality, increasing scientific noise and making it difficult to identify genuinely relevant evidence.[8] The challenge, therefore, lies not only in verifying whether references are genuine, but also in ensuring that the underlying science remains rigorous, clinically significant, and methodologically sound.

THE AI DETECTION TOOLS

This growing reality, marked by the rapid increase in the use of AI in the production of scientific literature, is amplified by the fact that it is not being accompanied by reliable methods and tools for accurately determining the extent of AI-generated content.

The scientific community's main concern is that low-quality or even completely fabricated articles may slip through current quality control systems, compromising the integrity of the scientific literature.[9]

As Miryam Naddaf concludes in her article recently published in Nature, there is clear evidence of a rapid increase in the use of AI in scientific output, but the exact extent of this phenomenon remains uncertain due to the limitations of current detection tools, namely their tendency to confuse text that has merely been edited by AI with text that is entirely AI-generated, and their occasional misclassification of human-written text as AI-generated.[9]

THE AUTOMATION OF PEER REVIEW AND THE EROSION OF JUDGMENT

A particularly concerning development in the scientific ecosystem is the delegation of peer review responsibilities to AI tools. Although many journals explicitly prohibit the use of AI for evaluating manuscripts due to risks of error, bias, and breaches of confidentiality, some reviewers have begun to use these systems to process their workloads.[2,5,6] Peer review is intended to be a process centered on human critical judgment. By transferring this responsibility to automated systems, the very purpose and value of peer review are hollowed out. When AI is used to evaluate manuscripts, the scientific community loses the rigorous human oversight necessary to detect subtle errors or systematic fraud. [2,5,6] Furthermore, LLMs have been found unsuitable for tasks such as detecting retracted literature due to issues with misinformation and false positives.[6]

CONCLUSION

The integration of AI into scientific publishing has moved beyond the theoretical and into the practical domain of daily editorial operations. The rise of hallucinated citations and the potential for a self-referential knowledge loop demand a reassertion of the human role in the scientific process. While AI can offer tools for efficiency, it

cannot replace the critical judgment, ethical responsibility, and rigorous verification that define scientific inquiry. Protecting the integrity of the scientific record requires a commitment to keeping the human element central to the circuit of knowledge production and validation. AI may accelerate scientific workflows, but it cannot replace the intellectual judgment required to determine whether scientific claims are valid, relevant, and worthy of becoming part of the permanent scientific record.

Ethical Disclosures

Conflicts of Interest: The authors have no conflicts of interest to declare.
Financial Support: This work has not received any contribution grant or scholarship.
Provenance and Peer Review: Commissioned; not externally peer-reviewed.
Consent for Publication: Not applicable.

REFERENCES

1. Ibrahim EI, Voyer A. Qualitative research with LLM chatbots: Technological reflexivity for interpretative technology. Qualitative Res. 2025;26:133-59. doi: 10.1177/14687941251390794
2. Donato H. Inteligência Artificial na Publicação Científica: Implicações para Autores, Revisores e Editores. Rev Port Med Interna. 2025;32:234-7.doi: 10.24950/rspmi.2846.
3. Naddaf M, Quill E. Hallucinated citations are polluting the scientific literature. What can be done? Nature. 2026;652:26-9. doi: 10.1038/d41586-026-00969-z.
4. Topaz M, Roguin N, Gupta P, Zhang Z, Peltonen LM. Fabricated citations: an audit across 2·5 million biomedical papers. Lancet. 2026;407:1779-81. doi: 10.1016/S0140-6736(26)00603-3.
5. Flanagin A, Kendall-Taylor J, Bibbins-Domingo K. Guidance for Authors, Peer Reviewers, and Editors on Use of AI, Language Models, and Chatbots. JAMA. 2023;330:702-3. doi: 10.1001/jama.2023.12500.
6. Metze K, Morandin-Reis RC, de Ávila Reis MF, da Silva Fago M, Florindo JB. Misinformation, false positives and delegation of tasks - Large Language Models should not be used for the detection of retracted literature - A study of 21 Chatbots. J Clin Anesth. 2025;107:112032. doi: 10.1016/j.jclinane.2025.112032.
7. Udesky L. 'Publish or perish' culture blamed for reproducibility crisis. Nature. 2025 (in press). doi: 10.1038/d41586-024-04253-w.
8. Naddaf M. Low-quality papers based on public health data are flooding the scientific literature. Nature. 2025 (in press). doi: 10.1038/d41586-025-02241-2.