

Drowning in Plausible Nonsense: Should We Implement AI Prechecks for Hallucinations and Misconduct in Editorial Processes?

Diego Borbón

Generative artificial intelligence (AI) tools hallucinate. They fabricate citations, invent references, and misattribute ideas to real or fake authors. Mathematical proofs suggest this is not a fixable bug but a structural feature of probabilistic language models. This essay argues that publishers should consider implementing automated prechecks designed to flag hallucinated citations, fabricated references, and misrepresented claims before manuscripts ever reach peer review. The tools to do so already exist and are increasingly operational. Without such prechecks, editors and reviewers are forced to spend their limited, frequently unpaid, time performing what amounts to *forensic auditing*; a task for which the editorial system was never designed and cannot sustain. There is an *asymmetry of negligence*: the less effort an author invests in producing a manuscript written by AI, the more effort it demands from everyone else to detect, document, and dispose of it. If the problem was created by the irresponsible use of generative AI (GenAI), the beginning of the solution could, paradoxically, also come from a better-informed use of AI.

Diego Borbón (<https://orcid.org/0000-0002-2115-2105>) is the Professor of Neuroscience, Artificial Intelligence, and Law at the Universidad Externado de Colombia and Visiting Researcher at the Fondation Brocher in Switzerland.

Opinions expressed are those of the authors and do not necessarily reflect the opinions or policies of their employers, the Council of Science Editors, or the Editorial Board of Science Editor.

<https://doi.org/10.36591/SE-4902-002>

Introduction

In 2023, with the boom of GenAI, Walters and Wilder¹ asked ChatGPT to produce literature reviews across 42 topics. Strikingly, 55% of citations generated by GPT-3.5 were entirely fabricated; for GPT-4 the figure was 18%.¹ Even among nonfabricated references, nearly a quarter contained substantive additional errors.

These are not marginal hallucination rates, as they describe the default behavior of probabilistic tools, now widely adopted by some to significantly boost their productivity and reduce the time it takes to produce a paper. For example, Kusumegi et al² found that scientists adopting large language models (LLMs) increased their publication output by 23.7%–89.3%. But, while the volume of AI-assisted submissions is rising, the reliability of their citational, substantive, and formal apparatus might not be.

This brief essay contends that editorial systems, which now receive significant amounts of manuscripts with documented hallucination rates, should consider implementing automated AI verification tools to precheck for hallucination in citation and references and to assess whether said citations misrepresent the cited authors' real claims and ideas. This would allow editors and reviewers to really focus on the main purpose of the scientific evaluation of manuscripts. And, for those papers that are flagged, following a human supervision of the automated results, editors could consider desk rejection instead of sending those manuscripts to reviewers, avoiding wasting the time doing otherwise.

This is not to say that publishers should adopt AI detectors and flag those that have a supposedly high probability of being AI generated. On the contrary, what the evidence suggests is that, first, AI detectors are not

CONTINUED

reliable, and are also biased against multilanguage authors, while ignoring the fact that a manuscript can be made with the help of AI and still comply with ethical, legal, and responsible standards and practices. What matters is not whether a manuscript was made with assistance of AI; if its responsible use shows an original input from the authors' ideas, a compelling and well-made argument, and has no hallucinations or misrepresentations, the use of AI can be accepted. What should matter then is whether the submission has profound errors in citations, attributions, and references to argue that it does not even grant the chance to be reviewed by others.

The Asymmetry of Negligence

We could call the *asymmetry of negligence* a structural condition that generative AI has introduced into academic publishing: The less effort an author invests in producing a manuscript written by AI, the more effort it demands from everyone else to detect, document, and dispose of it. An author can now generate a superficially competent and plausible manuscript in minutes, while conscientious reviewers, editors, and typesetters must then spend hours determining whether references, citations, and attributed/cited positions are real.

Frequently, an editor or a reviewer will receive a manuscript for which a significant amount of time would need to be dedicated for the labor of detecting false ideas, attributions, citations, and references, diminishing the time that can be dedicated to really assess what peer review is designed to evaluate: the quality of ideas and arguments; not checking for hallucinated mistakes. Then, generative AI has not merely lowered the barrier to writing plausible papers; it has obliterated it. And the editorial infrastructure built over centuries of voluntarism is buckling under the weight.

This is especially problematic, as Righi and Takács³ argued, because peer review constitutes a social dilemma in which noncontribution is the individually rational strategy. Bianchi et al⁴ showed that cooperation within peer review is extremely sensitive to external institutional conditions. But GenAI has radically altered those conditions. Before AI, the Publons⁵ report found that over 50% of review invitations were already declined, and 10% of reviewers performed half of all reviews. Aczel et al⁶ estimated unpaid review labor at over 100 million hours annually, worth more than US\$1.5 billion in the United States alone. This voluntary subsidy sustains an industry generating \$19 billion in revenue, with profit margins of 37%–40%.⁷ Asking these unpaid volunteers to now function almost as forensic auditors of AI-generated fabrication is a transfer of costs from those who profit and benefit from the system—including irresponsible authors—to those who fundamentally help sustain it.

The Feasibility Argument

The failure to implement automated citation-verification prechecks cannot be attributed to technological impossibility. For example, Scite.ai uses deep learning to classify citation statements as supporting, contrasting, or merely mentioning a cited work,⁸ currently indexing over 1.6 billion such statements.⁹ The STM Integrity Hub screens over 125,000 manuscripts monthly across 40 publishers, intercepting approximately 1,000 suspected paper mill submissions per month.¹⁰ Springer Nature has also deployed nearly 60 AI-powered tools across its editorial workflow, flagging 25,000 papers for potential integrity issues.¹¹ These are increasingly available tools, not just abstract hypothetical capabilities, anymore.

Limitations exist, of course, and they deserve honest acknowledgement. For example, the American Association of Cancer Research reported an analysis of 7,177 submitted manuscripts, finding that 36% of abstracts contained AI-generated text, but only 9% of authors disclosed its use.¹² The problem is that some of those analyses relied on an also unreliable (and sometimes off-point) tool: AI detectors.

Liang et al¹³ demonstrated that GPT-detection tools misclassify 61.3% of essays by multilanguage authors as AI generated. Hosseini and Resnik¹⁴ cautioned that current tools detect plagiarism of words but not misappropriation of ideas. These constraints do not justify inaction, as the use of AI detectors have, in my opinion, misguided AI's potential use in editorial processes.

What should matter for academic publishing is not whether someone used AI to assist research and writing, but whether its use compromises the basic requirements for science: citations need to truly represent someone's ideas, and the references included in a paper should genuinely represent the use of a real scientific precedent. Only then does it make sense to send a paper to a reviewer. If not, peer review would just turn into a complete waste of everybody's time.

When COPE first recommended the adoption of plagiarism-detection software, such as CrossCheck, the technology was also imperfect; its adoption was mandated not because it was infallible, but because it was available and superior to no screening at all. The same logic applies. Experienced reviewers rely on heuristic judgement, and those heuristics work when anomalies are exceptional. But when the base rate of citation fabrication and hallucinations are high, heuristic detection is overwhelmed: the cognitive system was not designed for an environment in which half of all references may just be complete plausible nonsense.

From a Logistical Gap to an Editorial Decision

The normative infrastructure for mandatory automated prechecks of hallucinations and AI misconduct already

exists, and the argument can proceed in three possible steps. The Committee on Publication Ethics Code (COPE) of Conduct and Best Practice Guidelines¹⁵ requires editors to take all reasonable steps to ensure the quality of the material they publish (§8.1), and recommends, as aspirational best practice, having systems in place to detect falsified data (e.g., inappropriately manipulated photographic images or plagiarised text), either for routine use or when suspicions are raised (§8, Best Practice). The Core Practices¹⁶ further requires journals to have clearly described processes for handling allegations of misconduct. If the availability of textual plagiarism detection generated an obligation to adopt it, then the availability of citation/reference/verification tools generates an analogous obligation. To accept one precedent and reject the other is an inconsistency, not a principled distinction.

Moreover, nowadays a regulatory asymmetry reveals the flaw. The National Institutes of Health¹⁷ and the Australian Research Council¹⁸ prohibit peer reviewers from using generative AI. And even some journals in *Nature*, *Science*, and *JAMA*, and some Taylor & Francis journals prohibit reviewers from entering manuscripts into chatbots—for justified reasons, of course. Yet Mollaki,¹⁹ as of 2024, found that most major publishers lacked specific policies addressing AI use by reviewers; although this picture has begun to evolve since, with several major publishers now issuing specific guidance for reviewers. This asymmetry, restrictions on reviewers while there is no corresponding obligation for editorial screening, is an institutional expression of the ethical omission.

The World Association of Medical Editors recommend that editors need appropriate detection tools, available regardless of ability to pay.²⁰ Without automated prechecks, the system delegates detection to reviewers and editors who generally lack a specific training for what can almost seem like a forensic auditing, and most have no access to verification platforms and, under open peer review, bear reputational consequences when undetected misconduct is published. This is a design flaw in editorial governance. The risk of automation bias²¹ is real, but it argues for human oversight of automated tools, not for their absence.

Scientific Integrity Under Threat

The stakes extend beyond the editorial workflow, and go even beyond what good science means, and what good faith entails. The academic citation/attribution/reference system is almost an act of epistemic trust. Cronin²² established that citing is a complex social practice embedded in recognition structures and knowledge claims. Nicolaisen²³ argued that citation behaviour is disciplined by a mechanism analogous to the handicap principle in evolutionary biology: authors avoid dishonest referencing because it threatens their reputation among knowledgeable peers.

Unfortunately, irresponsible use of generative AI dismantles this accountability mechanism and diminishes basic good faith and good science. Machines have no reputation to protect and no capacity for the intellectual engagement citation is meant to signify. When researchers have pointed out that between 14% and 92% of AI-generated citations are fabricated,²⁴ and when only 7% of AI-generated medical references are both authentic and accurate,²⁵ the citation becomes a simulacrum of trust; a structural ornament mimicking intellectual accountability while evacuating its substance. Citational integrity is a public good. Its erosion degrades the infrastructure of cumulative knowledge upon which all scholarship depends.

It is worth noting, in this regard, that the COPE position on authorship and AI tools is unambiguous: authors are fully responsible for the content of their manuscript, even those parts produced by an AI tool, and are thus liable for any breach of publication ethics.²⁶ The inclusion of fabricated references therefore already constitutes research misconduct and may lead to rejection or retraction, regardless of whether the fabrication was deliberate or inadvertent. GenAI makes inadvertent fabrication more likely, but it does not, and should not, change that outcome. Recent institutional responses are starting to reflect this principle. ArXiv, for example, has announced that authors who submit manuscripts containing hallucinated references face a 1-year ban from the platform, followed by a probationary requirement that subsequent submissions first be accepted at a peer-reviewed venue.²⁷ The signal is consistent with what is argued here: responsibility cannot be outsourced to the tool, and the failure to verify what one cites is not a minor lapse but a violation of the basic conditions of scholarly publication.

While it might be true that AI hallucinations might decrease with new and more robust models, the other crude truth is that it seems that fabrications are just a *natural* consequence of the probabilistic nature of GenAI. Kalai and Vempala²⁸ proved mathematically that any calibrated language model will inevitably hallucinate, since generating novel content from learned statistical distributions entails an irreducible probability of fabricating facts. Banerjee et al²⁹ extended this argument by showing that hallucinations are a structural property of LLMs rooted not only in probabilistic token generation but also in formal limitations, making them ineliminable regardless of architectural improvements or fact-checking mechanisms.

Then, it is almost safe to say that, at least for the near future, GenAI will hallucinate, and, as pointed out by Banerjee et al,²⁹ we need to live with this, but we should also do something about it, at least to not waste everyone's time in the editorial processes. Let's just think carefully what it means to not detect or precheck AI hallucinations in cited

CONTINUED

attributions and references. It means that either editors and reviewers will have to waste a significant amount of time reading the manuscript and then almost manually checking that citations and references make sense. Or it means that, if not detected, we are now publishing plausible but complete nonsense. Even if it passes through peer review, and typesetters then find hallucinations and errors in the citations and references after the paper was accepted, the solution cannot be to just correct the fabrications, as one cannot, in good conscience, conclude that the substantive content of the paper makes any sense, whatsoever.

Concluding Remarks

The question here is not whether fabricated citations will continue to enter the editorial process, but whether we choose to detect them before or after they waste everyone's time. The tools for automated citation and reference verification already exist, are already operational at scale, and follow the same logic that led the publishing community to adopt plagiarism detection software decades ago. The failure to deploy them is not a matter of technological readiness, but rather it is a matter of institutional decision.

Publishers who knowingly receive manuscripts with documented hallucination rates should consider the possibility of fighting irresponsible use of AI with responsible prechecks before transferring the costs of verification onto unpaid reviewers, overburdened editorial teams, and ultimately, the integrity of the published record itself. At the end, it is also in the best interest of the big publishers to avoid reputational costs, but also to keep the peer review system working. That starts by preventing the peer reviewers' work from becoming a waste of time and resources.

None of this implies that AI-assisted manuscripts should be rejected, flagged, or treated with suspicion simply because an author used a generative tool. What should concern us is not *just* how a paper was written, but whether its citations genuinely represent someone's ideas, and whether its references point to real scientific work. A manuscript that meets those basic conditions deserves to be reviewed on its merits, regardless of the tools used to produce it. A manuscript that does not meet them should never reach a reviewer's desk in the first place.

If the editorial system cannot reliably filter fraudulent and fabricated manuscripts, the entire architecture of science, that begins with the trust that one paper's citation of another represents a real engagement with real ideas, also begins to erode. The distinction is simple, the tools are available, and the cost of continued inaction is one the academic community can no longer afford. To avoid drowning in plausible nonsense, wasting everybody's time, it is time to think that maybe we should implement AI prechecks for hallucinations and misconduct in editorial processes.

Disclosures

The author declares no conflicts of interest. This essay was written during a research stay at the Fondation Brocher in Switzerland.

Resources and Links

1. Walters WH, Wilder EI. Fabrication and errors in the bibliographic citations generated by ChatGPT. *Sci Rep.* 2023;13:14045. <https://doi.org/10.1038/s41598-023-41032-5>
2. Kusumegi K, Yang X, Ginsparg P, de Vaan M, Stuart T, Yin Y. Scientific production in the era of large language models. *Science.* 2025;390:1240–1243. <https://doi.org/10.1126/science.adw3000>
3. Righi S, Takács K. The miracle of peer review and development in science: an agent-based model. *Scientometrics.* 2017;113:587–607. <https://doi.org/10.1007/s11192-017-2244-y>
4. Bianchi F, Grimaldo F, Bravo G, Squazzoni F. The peer review game: an agent-based model of scientists facing resource constraints and institutional pressures. *Scientometrics.* 2018;116:1401–1420. <https://doi.org/10.1007/s11192-018-2825-4>
5. Publons. Global state of peer review. Clarivate Analytics; 2018. <https://doi.org/10.14322/publons.GSPR2018>
6. Aczel B, Szaszi B, Holcombe AO. A billion-dollar donation: estimating the cost of researchers' time spent on peer review. *Res Integr Peer Rev.* 2021;6:14. <https://doi.org/10.1186/s41073-021-00118-2>
7. Hagve M. The money behind academic publishing. *Tidsskr Nor Laegeforen.* [published August 17, 2020; accessed June 11, 2026]. <https://tidsskriftet.no/en/2020/08/kronikk/money-behind-academic-publishing>
8. Nicholson JM, Mordaunt M, Lopez P, et al. scite: a smart citation index that displays the context of citations and classifies their intent using deep learning. *Quant Sci Stud.* 2021;2:882–898. https://doi.org/10.1162/qss_a_00146
9. <https://scite.ai/coverage>
10. STM Integrity Hub Team. The STM Integrity Hub: current status and recent developments. *Sci Ed.* 2025;48:167–168. <https://doi.org/10.36591/SE-4804-13>
11. Springer Nature Group. Springer Nature embraces AI tools across the publishing process. [published March 12, 2026; accessed June 11, 2026]. <https://group.springernature.com/la/group/media/press-releases/ai-tools-support-less-friction-and-increased-author-satisfaction/27849346>
12. Brainard J. Far more authors use AI to write science papers than admit it, publisher reports. *Science.* [published September 12, 2025; accessed June 11, 2026]. <https://www.science.org/content/article/far-more-authors-use-ai-write-science-papers-admit-it-publisher-reports>
13. Liang W, Yuksekgonul M, Mao Y, Wu E, Zou J. GPT detectors are biased against non-native English writers. *Patterns.* 2023;4:100779. <https://doi.org/10.1016/j.patter.2023.100779>
14. Hosseini M, Resnik DB. Guidance needed for using artificial intelligence to screen journal submissions for misconduct. *Res Ethics.* 2025;21:1–8. <https://doi.org/10.1177/17470161241254052>
15. Committee on Publication Ethics (COPE). Code of conduct and best practice guidelines for journal editors. [updated March 7, 2011; accessed June 11, 2026]. <https://publicationethics.org/media/609/download?attachment=>
16. Committee on Publication Ethics (COPE). COPE core practices. [accessed June 11, 2026]. <https://publication-ethics.org/resources/cope-core-practices/>

CONTINUED

17. National Institutes of Health. The use of generative artificial intelligence technologies is prohibited for the NIH peer review process, NOT-OD-23-149. [published June 23, 2023; accessed June 11, 2026]. <https://grants.nih.gov/grants/guide/notice-files/NOT-OD-23-149.html>
18. Australian Research Council. Confidentiality obligations of assessors. [published June 30, 2023; accessed June 11, 2026]. <https://www.arc.gov.au/news-publications/media/media-releases/confidentiality-obligations-assessors>
19. Mollaki V. Death of a reviewer or death of peer review integrity? The challenges of using AI tools in peer reviewing and the need to go beyond publishing policies. *Res Ethics*. 2024;20:239–250. <https://doi.org/10.1177/17470161231224552>
20. Zielinski C, Winker MA, Aggarwal R, et al. Chatbots, generative AI, and scholarly manuscripts: WAME recommendations. *Colomb Med*. 2023;54:e1015868. <https://doi.org/10.25100/cm.v54i3.5868>
21. Parasuraman R, Manzey DH. Complacency and bias in human use of automation: an attentional integration. *Hum Factors*. 2010;52:381–410. <https://doi.org/10.1177/0018720810376055>
22. Cronin B. The citation process: the role and significance of citations in scientific communication. Taylor Graham; 1984.
23. Nicolaisen J. Citation analysis. *Annu Rev Inf Sci Technol*. 2007;41:609–641. <https://doi.org/10.1002/aris.2007.1440401020>
24. Joubert M, Riedlinger M. Integrity under pressure: on generative AI, fabricated references and ethical publishing. *J Sci Commun*. 2026;25(1). https://jcom.sissa.it/article/pubid/JCOM_2501_2026_E/
25. Bhattacharyya M, Miller VM, Bhattacharyya D, Miller LE. High rates of fabricated and inaccurate references in ChatGPT-generated medical content. *Cureus*. 2023;15:e39238. <https://doi.org/10.7759/cureus.39238>
26. Committee on Publication Ethics (COPE). COPE position: authorship and AI tools. [reviewed February 13, 2023; June 11, 2026]. <https://publicationethics.org/guidance/cope-position/authorship-and-ai-tools>
27. Singh Chawla D. Researchers who use hallucinated references to face arXiv ban. *Nature*. 2026;653:988–989. <https://doi.org/10.1038/d41586-026-01595-5>
28. Kalai AT, Vempala SS. Calibrated language models must hallucinate. In: Proceedings of the 56th Annual ACM Symposium on Theory of Computing. ACM; 2024:160–171. <https://doi.org/10.1145/3618260.3649777>
29. Banerjee S, Agarwal A, Singla S. LLMs will always hallucinate, and we need to live with this. *arXiv*. 2024. <https://doi.org/10.48550/arXiv.2409.05746>