Peer Review in the Artificial Intelligence Era: A Call for Developing Responsible Integration Guidelines

Ahmed Salem BaHammam [iD][1−3]

[1]Editor-in-Chief Nature and Science of Sleep, [2]Department of Medicine, University Sleep Disorders Center and Pulmonary Service, King Saud University, Riyadh, Saudi Arabia; [3]King Saud University Medical City, Riyadh, Saudi Arabia

Correspondence: Ahmed Salem BaHammam, University Sleep Disorders Center, Department of Medicine, College of Medicine, King Saud University, Box 225503, Riyadh, 11324, Saudi Arabia, Tel +966-11-467-9495, Fax +966-11-467-9179, Email ashammam2@gmail.com

Introduction

The scientific publishing landscape has experienced a remarkable transformation, marked by an exponential surge in research output. Recent analyses reveal that scientific publications demonstrate an overall growth rate of 4.10% annually, with publications doubling approximately every 17.3 years.[1] This growth has intensified particularly in recent years, with medical research publishing and preprint servers placing unprecedented pressure on the peer review system.[2]

Several factors have contributed to this challenging situation. First, the pressure to "publish or perish" in academia has intensified, leading to more manuscript submissions across all disciplines.[3] Second, the emergence of new research fields and increasing specialization has created a demand for highly specific expertise among reviewers.[4]

The current peer-review system faces significant sustainability challenges, with a striking imbalance where 20% of researchers handle up to 94% of all reviews.[4] This inequitable distribution resulted in 63.4 million review hours in 2015 alone, with 18.9 million hours (30%) provided by just the top 5% of contributing reviewers.[4]

The combination of increasing submission volumes, growing specialization, and resource constraints has created bottlenecks in the review process. Journal editors often struggle to find qualified reviewers who can provide timely, high-quality feedback, leading to longer review times and potential delays in the dissemination of important research findings.

To mitigate this problem and address reviewer scarcity, several solutions have been proposed, including monetary compensation,[5] credits for future publications, providing access to the publisher's journals and databases,[6] and recognition through certificates, though payment remains controversial among publishers.[7,8] While proposed incentives like monetary compensation and publication credits appear promising, recent evidence reveals significant implementation hurdles. Financial rewards particularly burden smaller journals and those in developing nations,[7] while certificates and publication credits have shown limited effectiveness in boosting reviewer engagement.[8]

Against this backdrop, the emergence of Large Language Models (LLMs) represents a potential turning point in academic publishing. These sophisticated AI systems, exemplified by tools like ChatGPT, have sparked intense discussion within the academic community about their potential role in scientific writing[9] and, more recently, in the peer review processes.[10,11] Given the mounting pressures on the traditional peer-review system and the rising pressing questions about system sustainability, especially as submission volumes continue to climb, the growing interest in AI-assisted peer review is unsurprising.

As concerns about AI's role in scientific writing continue to grow, its potential application in peer review raises even more fundamental questions about the integrity and confidentiality of manuscript evaluation.[12] While there is no systematic evidence quantifying actual usage, anecdotal reports suggest some reviewers may be considering AI assistance despite institutional and publishers' prohibitions.[13] The demanding nature of peer review work, combined with minimal academic recognition, could create incentives for seeking AI-enabled efficiency.

Despite the potential benefits of AI assistance in scholarly publishing, the academic community remains divided. Academic publishers vary in their approach to AI in peer review, with some lacking formal policies, others allowing

159

limited use, and some outright prohibiting it. A recent analysis of top medical journals revealed that 78% now provide specific guidance on AI use in peer review, with 59% of them explicitly prohibiting its use.[2] This divergence in policy reflects the broader tension between innovation and tradition in academic publishing.

This editorial addresses the critical and timely challenge of integrating AI into peer review, a rapidly emerging issue poised to reshape scholarly publishing. It examines the potential benefits, ethical dilemmas, and the pressing need for guidelines to balance innovation with research integrity.

Potential Benefits and Capabilities in Peer Review

The idea of integrating AI into peer review presents a promising development in scholarly publishing. Recent large-scale analyses suggest potential benefits of AI-assisted review processes, offering possible solutions to address the growing challenges in academic publishing.

Recent studies analyzing over 3000 Nature family journal papers and 1700 computer science conference papers show that AI-generated reviews achieve specific overlap rates with human reviewer comments: 30.85% for Nature family journals and 39.23% for ICLR (International Conference on Learning Representations) papers.[14] While Saad et al found moderate alignment between AI and human reviews (mean scores of 3.6/5 for ChatGPT 3.5 and 3.76/5 for ChatGPT 4.0), AI tends to be better at identifying technical and methodological issues than evaluating scientific impact or novelty.[15] In fact, AI is 10.69 times less likely to comment on novelty compared to human reviewers.[14] This notable limitation in evaluating novelty is particularly concerning since assessing the originality and innovative aspects of research is a fundamental aspect of peer review. This limitation becomes especially problematic when assessing interdisciplinary research requiring domain expertise and scientific intuition.

Recent studies demonstrate that AI tools can flag obvious protocol deviations and experimental design flaws.[16] However, they struggle with assessing the appropriateness of specialized experimental methods, particularly in novel or complex research designs. The assessment of experimental rigor and methodological appropriateness still requires human expertise and judgment.

A survey of 308 researchers from 110 US institutions in the domain of AI and computational biology revealed that while AI-generated feedback aligned with standard review expectations, it lacked the specificity of human reviews. Although half would use AI tools again, most preferred traditional peer review,[14] highlighting AI's role as a complementary tool rather than a replacement for human expertise. Moreover, AI systems have the potential to contribute unique perspectives to the review process. Over 65% of researchers report that AI feedback identifies points overlooked by human reviewers, while more than 57% find AI-generated feedback helpful in improving their work before formal submission.[14] This complementary insight proves particularly valuable during manuscript preparation, where early AI feedback can significantly improve paper quality.

The current peer review system demands approximately 100 million researcher hours and $2.5 billion annually.[14] AI systems significantly reduce this burden through rapid initial screening and automated technical verification. These systems can potentially evaluate manuscript compliance, verify formatting requirements, and identify methodological inconsistencies within minutes rather than hours.[11]

The UK's Research Excellence Framework 2021 cost approximately £471 million, with institutional preparation costs accounting for £430 million of this total.[17] While administrative processes like data collection and assessment consume significant resources, AI tools could potentially assist with technical verification and initial screening tasks, complementing other efficiency measures being implemented in the review process.[17,18]

Practical Applications in Peer Review

AI tools have the potential to serve three distinct functions in manuscript evaluation. First, they could help in screening and analyzing large volumes of scientific literature, which is particularly valuable for systematic reviews requiring the assessment of thousands of papers.[19,20] Second, early evidence suggests that AI systems may assist with administrative tasks like reviewer selection. The National Health and Medical Research Council (Australia) already uses AI to match grant proposals with suitable reviewers, though this remains limited to administrative support rather than quality

evaluation.[18] Third, they excel at technical verification tasks like checking statistical analyses, reference accuracy, and formatting requirements.[15,21]

ChatGPT models, while initially trained on historical data, now incorporate real-time search capabilities through plugins and integrations, allowing access to the latest peer-reviewed research and ensuring decisions are informed by current evidence, though careful implementation and human oversight remain critical.[22] Nevertheless, it is necessary to validate the timeliness and accuracy of information cited or analyzed by AI models, ensuring their responsible and effective integration into academic workflows.

While AI tools show promise in enhancing the peer review process, their role in statistical verification remains constrained by several key limitations that require further clarification. Current AI systems can assist with technical verification in several specific ways. First, they help identify inconsistencies between statistical methods described in the methods section and those reported in the results.[16] For example, they can flag when a paper claims to use ANOVA in methods but reports only t-test results or when effect sizes are mentioned but not actually reported. Recent studies show GPT-4 achieves 74% accuracy in identifying such statistical inconsistencies across diverse datasets.[16] Second, these tools are effective at detecting missing essential statistical information according to reporting guidelines.[15] This includes unreported p-values, missing confidence intervals, undefined significance thresholds, or incomplete descriptive statistics. For instance, GPT-4 can identify when means are reported without standard deviations or when test statistics lack degrees of freedom or remind the reviewer that sample size calculation was not reported. AI tools demonstrate particular strength in text normalization and classification tasks related to statistical reporting elements.[16] Third, AI can flag potential discrepancies, such as when conclusions do not align with the reported significance levels or when statistical terminology is used incorrectly. This capability is particularly evident in inferential statistics and sampling theory, where AI shows strong performance in identifying methodological issues.[16]

Beyond these technical capabilities, research has also examined AI's broader impact on the peer review process. Recent preliminary research has demonstrated that AI-assisted reviews can significantly impact paper evaluation outcomes, with submissions receiving AI-assisted reviews being 4.9 percentage points more likely to be accepted[23]—a controversial benefit raising concerns about potential biases. However, this performance varies considerably depending on the complexity and structure of the statistical questions, with accuracy ranging from moderate to high across different statistical domains.[16]

While the above capabilities and impacts are promising, it is important to note several key limitations regarding statistics assessment: AI tools cannot validate the appropriateness of chosen statistical approaches or verify actual calculations without access to raw data.[15,18] Recent research shows that while AI reviewers can provide evaluations of scientific work, they tend to give inflated quality ratings and fail to distinguish between high- and low-quality research output.[24]

Critical Concerns in the Quality of AI-Assisted Peer Review

A compelling demonstration of this weakness was reflected in Thelwall's experiment, where he altered a legitimate research paper about gender differences in surgeon citation impact to focus on a fictitious comparison. He replaced "male" with "squirrel" and "female" with "human" throughout the text, transforming the study into an absurd analysis titled, "Do squirrel surgeons generate more citation impact?[25]" Despite the nonsensical premise of comparing 'squirrel' and 'human' surgeons, the AI system rated the paper highly and failed to identify its fundamental flaws.[18,25] AI systems also struggle with evaluating more nuanced aspects, like whether sample sizes are adequate or if statistical assumptions are met. These limitations emphasize the critical need for human supervision in AI-powered peer review to ensure accuracy and maintain scientific rigor.[11] While the agreement between human reviewers themselves is often low, showing only about 20% agreement in quality assessments, the agreement between human and AI reviewers is similarly low but becomes stronger when evaluating high-quality research.[24] Recent literature strongly suggests that while AI can enhance the efficiency of peer review, it should function as a complementary tool rather than a replacement for expert judgment.[23]

These limitations extend beyond identifying absurd content. While AI systems generate plausible-sounding reviews, they often provide superficial feedback that lacks the nuanced critique necessary for advancing scientific discourse.[15] The AI-generated evaluations struggle particularly with established assessment frameworks, showing accuracy rates below

acceptable thresholds for high-stakes academic decisions.[25] The core challenge lies in the inability of AI to replicate human expert judgment.

Confidentiality Concerns in AI-Assisted Peer Review

Among the various challenges facing AI integration in peer review, confidentiality emerges as the most critical concern, extending beyond traditional data privacy issues to potentially compromising the foundational trust between authors, reviewers, and publishers. When manuscripts are uploaded to public AI (LLMs) platforms for review assistance, this constitutes a direct breach of confidentiality agreements that form the foundation of the peer review process.[13] The National Institute of Health (NIH) explicitly prohibits scientific peer reviewers from using natural language processors, LLMs, or other generative AI technologies for analyzing and formulating peer review critiques, as this would violate the confidential nature of the review process.[13]

The security of proprietary research data presents another critical challenge. Commercial AI platforms provide no guarantees regarding data retention or usage, potentially storing submitted manuscripts indefinitely and using them for model training.[26] This creates significant risks for authors, particularly when manuscripts contain sensitive intellectual property or unpublished discoveries. Recent regulatory guidelines emphasize that sharing unpublished research through AI platforms poses substantial risks to both data privacy and intellectual property protection.[27]

Future Directions and Implementation Strategy

The exponential growth in scientific publications, combined with mounting financial pressures and reviewer fatigue, makes AI an increasingly appealing solution for peer review challenges. This substantial investment of time and resources, coupled with a growing shortage of qualified reviewers, has created an urgent need to explore innovative solutions.

The adoption of AI in peer review represents a critical inflection point in academic publishing, requiring careful consideration of both technological capabilities and ethical imperatives. Implementation could begin with AI integration in internal evaluation phases, where AI tools can identify basic errors and technical issues, allowing human editors to focus on manuscript content and scientific merit. This initial screening process must maintain human oversight of AI-generated feedback to ensure accuracy and prevent bias.[28]

The scientific community faces several pressing challenges, including determining the optimal balance between AI assistance and human expertise, evaluating the impact on review quality, and developing secure, discipline-specific solutions. Clear guidelines for joint human-AI review are essential, with human editors verifying AI feedback for inaccuracies or bias. This mutual supervision approach helps prevent algorithmic errors while maximizing the efficiency gains of AI assistance.[28] A key challenge lies in upgrading legacy review systems while maintaining data security and addressing resistance to automation.[29] Equally important is developing standardized frameworks for assessing AI-assisted review quality and preventing misuse.[11]

Success requires coordinated participation from multiple stakeholders. Publishers may integrate AI tools within their internal review processes, supporting human editors by identifying basic errors before manuscripts reach external peer reviewers. These AI systems should undergo regular retesting and retraining to improve validity and account for field advances.[28] Equal attention must be paid to developing comprehensive training programs for editors and reviewers, particularly in addressing bias and maintaining scientific discourse quality.[28] Publishers must establish secure systems that protect manuscript confidentiality throughout the peer review process, with clear policies prohibiting the upload of unpublished manuscripts to public AI platforms.[13] Any integration of AI tools should occur within protected journal management systems that maintain strict data privacy standards. Organizations successfully integrating AI typically follow a hybrid approach emphasizing continuous training, clear ethical guidelines, and technological advancement.[29] This careful balance between innovation and tradition ensures that technological advancement enhances rather than compromises peer review integrity.

As AI integration in scholarly publishing becomes increasingly inevitable, the academic community must proactively shape its implementation rather than merely react to its emergence. This editorial serves as a call to action for all stakeholders to engage in meaningful dialogue about responsible AI integration in peer review. Through continued

collaboration and careful consideration of ethical implications, we can work toward maximizing AI's benefits while preserving peer review integrity.

Nature and Science of Sleep maintains strict guidelines protecting manuscript confidentiality by prohibiting unauthorized AI use in peer review, while requiring disclosure of any permitted AI assistance and will regularly update policies as the technology evolves.[30]

Author Contributions

Ahmed S. BaHammam: Conceptualization, Writing – original draft preparation, Writing – review and editing, and Supervision. Ahmed S. BaHammam agreed to the journal to which the article will be submitted and agreed to take responsibility and be accountable for the article's contents.

Funding

Disclosure

The author reports no conflicts of interest in this work.

References

1. Bornmann L, Haunschild R, Mutz R. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*. 2021;8(1):224. doi:10.1057/s41599-021-00903-w
2. Li ZQ, Xu HL, Cao HJ, Liu ZL, Fei YT, Liu JP. Use of Artificial Intelligence in Peer Review Among Top 100 Medical Journals. *JAMA Netw Open*. 2024;7(12):e2448609. doi:10.1001/jamanetworkopen.2024.48609
3. Guraya SY, Norman RI, Khoshhal KI, Guraya SS, Forgione A. Publish or Perish mantra in the medical field: a systematic review of the reasons, consequences and remedies. *Pak J Med Sci*. 2016;32(6):1562–1567. doi:10.12669/pjms.326.10490
4. Kovanis M, Porcher R, Ravaud P, Trinquart L. The Global Burden of Journal Peer Review in the Biomedical Literature: strong Imbalance in the Collective Enterprise. *PLoS One*. 2016;11:e0166387.
5. Seghier ML. Paying reviewers and regulating the number of papers may help fix the peer-review process. *F1000Res*. 2024;13:439. doi:10.12688/f1000research.148985.1
6. Gasparyan A, Gerasimov A, Voronov A, Kitas G. Rewarding Peer Reviewers: maintaining the Integrity of Science Communication. *Journal of Korean Medical Science*. 2015;30(4):360–364. doi:10.3346/jkms.2015.30.4.360
7. Belém de Oliveira Neto J. The challenge of reviewers scarcity in academic journals: payment as a viable solution. *Einstein*. 2024;22:eED1194. doi:10.31744/einstein_journal/2024ED1194
8. Henderson S, Berk M, Boyce P, et al. Finding reviewers: a crisis for journals and their authors. *Australian & New Zealand Journal of Psychiatry*. 2020;54(10):957–959. doi:10.1177/0004867420958077
9. BaHammam AS. Balancing Innovation and Integrity: the Role of AI in Research and Scientific Writing. *Nat Sci Sleep*. 2023;15:1153–1156. doi:10.2147/NSS.S455765
10. Biswas S, Dobaria D, Cohen HL. ChatGPT and the Future of Journal Reviews: a Feasibility Study. *Yale J Biol Med*. 2023;96(3):415–420. doi:10.59249/SKDH9286
11. Hosseini M, Horbach S. Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. *Res Integr Peer Rev*. 2023;8(1):4. doi:10.1186/s41073-023-00133-5
12. Bahammam AS, Trabelsi K, Pandi-Perumal SR, Jahrami H. Adapting to the Impact of Artificial Intelligence in Scientific Writing: balancing Benefits and Drawbacks while Developing Policies and Regulations. *Journal of Nature and Science of Medicine*. 2023;6(3):152–158. doi:10.4103/jnsm.jnsm_89_23
13. Lauer M, Constant S, Wernimont A. Using AI in Peer Review Is a Breach of Confidentiality. 2023. Accessed December 18, 2024. Available from: https://nexus.od.nih.gov/all/3/06/23/using-ai-in-peer-review-is-a-breach-of-confidentiality/.
14. Liang W, Zhang Y, Cao H, et al. Can Large Language Models Provide Useful Feedback on Research Papers? A Large-Scale Empirical Analysis. *NEJM AI*. 2024;1(8). doi:10.1056/AIoa2400196
15. Saad A, Jenko N, Ariyaratne S, et al. Exploring the potential of ChatGPT in the peer review process: an observational study. *Diabetes Metab Syndr*. 2024;18(2):102946. doi:10.1016/j.dsx.2024.102946
16. Costa D, Dias J, Nakano C, Monteiro ML. Assessing the Statistical Abilities of GPT-4 as a Zero-shot Reasoner: strengths and Weaknesses. *Research Square*. 2023;2023:1–17. doi:10.21203/rs.3.rs-2958780/v1
17. UK Research and Innovation. New Future Research Assessment Programme reports. 2023. Accessed December 12, 2024. Available from: https://www.ukri.org/news/new-future-research-assessment-programme-reports-published/.
18. Ryan J. Must peer reviewers be human to assess quality? *Nature*. 2024;633(Suppl 1):S18–S20. doi:10.1038/633.S18
19. Buetow S, Lovatt J. From insight to innovation: harnessing artificial intelligence for dynamic literature reviews. *The Journal of Academic Librarianship*. 2024;50(4):102901. doi:10.1016/j.acalib.2024.102901
20. Feng Y, Liang S, Zhang Y, et al. Automated medical literature screening using artificial intelligence: a systematic review and meta-analysis. *Journal of the American Medical Informatics Association: JAMIA*. 2022;29(8):1425–1432. doi:10.1093/jamia/ocac066

21. D'Ambrosio A, Grundmann H, Donker T. An open-source integrated framework for the automation of citation collection and screening in systematic reviews. *ArXiv.* 2022. abs/2202.10033.

22. Yip R, Sun YJ, Bassuk A, Mahajan VB. Artificial Intelligence's Contribution to Biomedical Literature Search: revolutionizing or Complicating? *bioRxiv.* 2024;2024.10.07.617112.

23. Latona GR, Ribeiro MH, Davidson TR, Veselovsky V, West R. The AI Review Lottery: widespread AI-Assisted Peer Reviews Boost Paper Scores and Acceptance Rates.. *arXiv.* 2024. arXiv:2405.02150. doi:10.48550/arXiv.2405.02150.

24. Shcherbiak A, Habibnia H, Böhm R, Fiedler S. Evaluating science: a comparison of human and AI reviewers. *Judgment and Decision Making.* 2024;19. doi:10.1017/jdm.2024.24

25. Thelwall M. Can ChatGPT evaluate research quality? *Journal of Data and Information Science.* 2024;9(2):1–21. doi:10.2478/jdis-2024-0013

26. Strowel A. ChatGPT and Generative AI Tools: theft of Intellectual Labor? *IIC International Review of Intellectual Property and Competition Law.* 2023;54(4):491–494.

27. SDAIA. Generative Artificial Intelligence Guidelines Public. 2024. Accessed December 1, 2024. Available from: https://sdaia.gov.sa/en/SDAIA/about/Files/GenAIGuidelinesForGovernmentENCompressed.pdf.

28. Sabet CJ, Bajaj SS, Stanford FC, Celi LA. Equity in scientific publishing: can artificial intelligence transform the peer review process? *Mayo Clinic Proceedings: Digital Health.* 2023;1(4):596–600. doi:10.1016/j.mcpdig.2023.10.002

29. 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. *Research Ethics.* 2024;20(2):239–250. doi:10.1177/17470161231224552

30. Taylor & Francis. AI Policy: Taylor & Francis; 2024. Accessed Jan 21, 2025. Available from: https://taylorandfrancis.com/our-policies/ai-policy/.