ORIGINAL ARTICLE

(wileyonlinelibrary.com) doi: 10.1002/leap.1638

Received: 25 July 2024 | Accepted: 1 October 2024

Enhancing peer review efficiency: A mixed-methods analysis of artificial intelligence-assisted reviewer selection across academic disciplines

Shai Farber

Department of Law and Criminology, Emek Jezreel Academic College, Artificial Intelligence Researcher, Jezreel Valley, Israel

ORCID: 0000-0002-6252-805X

E-mail: shaif@yvc.ac.il

Abstract: This mixed-methods study evaluates the efficacy of artificial intelligence (AI)-assisted reviewer selection in academic publishing across diverse disciplines. Twenty journal editors assessed AI-generated reviewer recommendations for a manuscript. The AI system achieved a 42% overlap with editors' selections and demonstrated a significant improvement in time efficiency, reducing selection time by 73%. Editors found that 37% of AI-suggested reviewers who were not part of their initial selection were indeed suitable. The system's performance varied across disciplines, with higher accuracy in STEM fields (Cohen's $d = 0.68$). Qualitative feedback revealed an appreciation for the AI's ability to identify lesser-known experts but concerns about its grasp of interdisciplinary work. Ethical considerations, including potential algorithmic bias and privacy issues, were highlighted. The study concludes that while AI shows promise in enhancing reviewer selection efficiency and broadening the reviewer pool, it requires human oversight to address limitations in understanding nuanced disciplinary contexts. Future research should focus on larger-scale longitudinal studies and developing ethical frameworks for AI integration in peer-review processes.

Keywords: academic publishing, artificial intelligence, editorial efficiency, peer review, reviewer selection

INTRODUCTION

Background on peer review in academic publishing

Peer review, the cornerstone of academic publishing, has been integral to the advancement of scientific knowledge for over three centuries (Csiszar, 2016). This process, whereby experts in a given field critically evaluate the work of their peers before publication, serves as a crucial quality control mechanism in scholarly communication (Tennant et al., 2017). While peer review has stood the test of time, it continues to evolve dynamically, adapting to the shifting landscape of academic research and technological progress (Horbach & Halffman, 2018).

One of the most persistent challenges in the peer-review process is the selection of appropriate reviewers. Editors face the task of identifying experts who possess not only the requisite knowledge but also the impartiality and availability to provide timely, constructive feedback (Kovanis et al., 2016). The exponential growth in research output, coupled with the increasing specialization of academic fields, has exacerbated this challenge (Publons, 2018). Moreover, modern academia's global nature necessitates considering diverse perspectives, further complicating the reviewer selection process (Rodríguez-Bravo et al., 2017). Despite its pivotal role in academic publishing, it is crucial to acknowledge the limitations of peer review. Godlee and Jefferson's (1999) comprehensive analysis highlights that the process is not infallible: it can be slow, may overlook certain methodological flaws, and is susceptible to various biases. Recognizing these constraints is essential as we investigate potential artificial intelligence (AI)-assisted enhancements to the peer-review process.

AI has emerged as a transformative force across various sectors, and academic publishing is no exception. The potential of AI to revolutionize scholarly processes, from manuscript submission to publication, has garnered significant attention in recent years (Teixeira da Silva et al., 2019). Notably, the application of machine learning algorithms to analyse vast datasets of academic literature and researcher profiles offers promising avenues for enhancing the efficiency and effectiveness of peer review (Heaven, 2018).

AI-driven tools have already demonstrated their capacity to assist in plagiarism detection, reference checking, and even basic language editing. For instance, tools like Turnitin and Grammarly have integrated AI to assist with plagiarism detection and language editing, respectively, showcasing practical applications of AI in academic settings. Such tools have become invaluable resources for both students and researchers, improving the quality and integrity of academic writing.

This study aims to evaluate the efficacy of AI-assisted reviewer selection in academic publishing. By comparing AI-generated recommendations with traditional selection methods employed by experienced journal editors, the study seeks to:

1. Assess the accuracy and relevance of AI-suggested reviewers across diverse academic disciplines.
2. Quantify potential time savings and efficiency gains in the reviewer selection process.
3. Explore editors' perceptions and attitudes towards AI integration in peer review.
4. Identify potential ethical considerations and limitations of AI-assisted reviewer selection.

The significance of this research lies in its potential to address a critical bottleneck in the academic publishing process. As the volume of submissions continues to grow, innovative solutions are needed to maintain the quality and timeliness of peer review (Arns, 2014). By rigorously evaluating the performance of AI in reviewer selection, this study contributes to the ongoing dialogue on the future of peer review and the responsible integration of AI in scholarly communication. The findings may inform best practices for journal editors and publishers, potentially leading to more efficient, equitable, and effective peer-review processes in the digital age.

Key points

- The study found that AI reduced reviewer selection time by 73%, offering significant time savings for journal editors, especially in STEM fields where it demonstrated higher accuracy (52% overlap with editors' selections).
- AI systems helped identify 37% of suitable reviewers not initially considered by editors, with particular success in suggesting lesser-known experts, potentially increasing diversity in the peer review process.
- AI showed lower accuracy in social sciences and interdisciplinary fields (35% overlap), often missing methodological nuances or proposing overly senior reviewers.
- Editors raised concerns about algorithmic bias, geographic and institutional disparities, and instances of AI suggesting fictional reviewers (10% of cases), highlighting the need for human oversight.
- The study emphasizes the importance of developing AI systems customized for different academic fields to address disciplinary variations in performance and reviewer suitability.

LITERATURE REVIEW

Traditional methods of reviewer selection

Selecting appropriate peer reviewers has long been a critical and challenging aspect of academic publishing. Traditionally, editors have relied on a combination of personal knowledge, professional networks, and manual searches of relevant literature to identify suitable reviewers (Mulligan et al., 2013). While often effective, this traditional approach is typically time-consuming and can be constrained by the editor's own expertise and connections within the field (Kovanis et al., 2017).

Journal-specific reviewer databases have been a common tool, allowing editors to track reviewer performance and specialties over time (Gasparyan et al., 2015). However, these databases require constant updating and may not adequately capture the full spectrum of potential reviewers, particularly in rapidly evolving or interdisciplinary fields (Severin & Chataway, 2021).

In recent years, some journals have implemented open calls for reviewers or author-suggested reviewer systems. While these methods can broaden the pool of potential reviewers, they also introduce new challenges, including potential conflicts of interest or the risk of suggested reviewers being fabricated or coerced (Rivera, 2018).

AI applications in academic publishing

AI has started to infiltrate various aspects of academic publishing, offering promising solutions to long-standing challenges. In the context of peer review, AI applications span from automated plagiarism detection to more complex tasks such as matching manuscripts with suitable reviewers.

Natural language processing (NLP) techniques are being utilized to analyse manuscript content and compare it with potential reviewers' expertise, as demonstrated by their publication history (Mrowinski et al., 2017). These AI-driven systems aim to identify reviewers whose research interests closely align with the subject matter of the submitted manuscript, potentially improving the quality and relevance of peer review.

Machine learning algorithms have also been developed to predict reviewer performance based on factors such as past review quality, timeliness, and citation impact (Price & Flach, 2017). Such predictive models could assist editors in selecting reviewers who are not only knowledgeable but also likely to provide timely and constructive feedback. However, current AI applications in academic publishing face several limitations. These include:

1. Limited contextual understanding: AI systems may struggle to grasp nuanced or interdisciplinary research topics.
2. Potential for bias: AI models trained on historical data may perpetuate existing academic biases.
3. Lack of transparency: The 'black box' nature of some AI algorithms can make it difficult to understand how decisions are made.
4. Difficulty assessing soft skills: AI may not effectively evaluate important reviewer qualities such as critical thinking or constructive criticism abilities.

It is important to acknowledge that several organizations are already addressing some of the limitations identified in this study through specialized AI-powered reviewer selection tools. Examples include Prophy.ai's Referee Finder, Global Campus AI, and Dimensions.ai's Reviewer Finder. These tools often employ fine-tuned models and specialized algorithms that may offer improved performance in specific disciplines or contexts. While our study focused on a general AI model, the existence and development of these specialized tools highlight the dynamic nature of this field and the ongoing efforts to enhance AI-assisted reviewer selection.

Some researchers envision a future where AI's role in academic publishing extends far beyond these current applications. Habibzadeh (2023) suggests that AI could eventually handle most aspects of research and publication, from data analysis to article generation, potentially eliminating the need for traditional journals and peer-review processes. Such predictions highlight the need for ongoing discussion about the long-term implications of AI in scholarly communication.

Ethical considerations in AI-assisted peer review

The incorporation of AI into the peer-review process raises significant ethical considerations. A primary issue is the potential for algorithmic bias, where AI systems may perpetuate or even amplify existing biases in academia (Rodríguez-Bravo et al., 2017). For example, if historical data used to train AI models reflects gender or geographical disparities in academic publishing, these biases could be reinforced in AI-assisted reviewer selection. Another significant ethical consideration is the transparency and explainability of AI decision-making processes. The 'black box' nature of some machine learning algorithms can make it difficult to understand how reviewer recommendations are generated, potentially undermining trust in the peer-review process (Horbach & Halffman, 2020).

Privacy and data protection concerns also come to the fore when considering the vast amounts of personal and professional data that AI systems may process to generate reviewer recommendations (Teixeira da Silva et al., 2019). Ensuring compliance with data protection regulations and maintaining the confidentiality of both authors and potential reviewers is crucial.

Current research

Despite the growing interest in AI applications for peer review, several significant gaps remain in the current research landscape. First, there is a paucity of large-scale, empirical studies comparing the effectiveness of AI-assisted reviewer selection with traditional methods across diverse academic disciplines (Heaven, 2018). Second, the long-term impacts of AI integration on the quality and integrity of peer review remain underexplored. While AI promises increased efficiency, its effects on the depth, diversity, and constructiveness of peer feedback are yet to be fully elucidated (Tennant et al., 2019).

Third, there is limited research on the perceptions and attitudes of key stakeholders—including editors, reviewers, and authors—towards AI-assisted peer review. Understanding these perspectives is crucial for successfully implementing and accepting AI tools in academic publishing (Severin et al., 2020). Finally, the development of ethical frameworks and best practices for the responsible use of AI in peer review is still in its infancy. Additional research is imperative to establish guidelines that balance the potential benefits of AI with the fundamental principles of fairness, transparency, and academic integrity (Grimaldo et al., 2018).

In light of the above, this study addresses three primary research questions:

TABLE 1 Distribution of participating editors by academic discipline.

Discipline	Number of editors
Law	2
Psychology	2
Sociology	2
Education	2
Electrical Engineering	2
Medicine	2
Biology	2
Public Health	2
Economics	2
Chemistry	2

1. To what extent does AI-assisted reviewer selection improve efficiency compared with traditional methods across different academic disciplines?
2. How accurate are AI-generated reviewer recommendations compared to expert editor selections?
3. What are the key ethical considerations and potential biases in implementing AI-assisted reviewer selection?

These research questions aim to comprehensively evaluate AI's potential in peer review, addressing quantitative efficiency, accuracy, and qualitative ethical concerns. This study fills a significant research gap by providing empirical evidence on AI-assisted reviewer selection across diverse disciplines, comparing it with traditional methods and examining editors' perceptions. The mixed-methods approach offers insights into both quantitative gains and qualitative implications of AI integration in peer review.

METHODOLOGY

Research design

This study utilized a mixed-methods approach to evaluate the efficacy of AI-assisted reviewer selection in academic publishing. By integrating both quantitative and qualitative elements, the research design facilitated a comprehensive assessment of the AI system's performance compared to traditional reviewer selection methods.

Participant selection

Twenty editors, each representing a different academic journal, were recruited to participate in the study through a snowball sampling method. Initial contact was made with a broad range of editors, some of whom were recruited based on personal acquaintance. These editors then referred their colleagues to participate. This approach ensured a diverse range of disciplines, as outlined in Table 1. The gender distribution of the participants was 60% female and 40% male, with an average age of 54 years. All participants had at least 3 years of experience in their editorial roles, with some having up to 15 years of experience. The average editorial experience in the current sample was 7.1 years. All participants provided informed consent prior to their involvement in the study.

AI system

This study employed an AI system based on the commercially available GPT-4 model, accessed through its paid subscription service. Journal editors employed this advanced language model to assist in the reviewer selection process for academic manuscripts. The GPT-4 model, which has been trained on a vast corpus of text, including academic literature, was used to analyse manuscript content and suggest potential reviewers based on relevant expertise and research background (further details on the model used can be found in Appendix C). Editors input manuscript details into the GPT-4 interface, which then generated recommendations for suitable peer reviewers. This approach leveraged the model's ability to process and understand complex academic text, enabling it to identify potential matches between manuscript content and researcher expertise across diverse disciplines.

To address confidentiality concerns, only the title, abstract, and keywords of each manuscript were provided to the AI system without any author information. We obtained consent from journal editors to use this limited manuscript information for the study. No full manuscripts were uploaded to the AI system to maintain confidentiality and intellectual property rights. The full prompt used to instruct the AI system is provided in Appendix A. It is important to note that the use of ChatGPT in this study was strictly for research purposes, and we adhered to ethical guidelines regarding data usage and privacy.

The GPT-4 model was accessed independently by each of the 20 participating editors through the ChatGPT4 Pro subscription service. All inputs were entered during the week of 22–29 April 2024. This timeframe was chosen to ensure that all editors were using the same version of the model, as GPT-4 undergoes continuous updates. It is important to note that while efforts were made to standardize the timing, the dynamic nature of AI models means that minor variations in responses might occur even within this short timeframe. This approach allows for a more realistic assessment of how the AI system would perform in actual editorial practice, where multiple users might access the system independently over a period of time.

While the AI prompt included instructions to avoid suggesting reviewers with potential conflicts of interest, we acknowledge that author names were not included in the input. This limitation means that the AI system's ability to identify conflicts of interest was restricted. Editors were instructed to review the AI suggestions carefully and apply their knowledge of potential conflicts when making final selections.

Data collection procedure

Each participating editor was asked to follow a standardized procedure:

TABLE 2 Summary of key performance metrics.

Metric	Overall performance	STEM (Science, Technology, Engineering, and Mathematics)	Social Sciences & Humanities
Selection accuracy	42% (95% CI: 39%–45%)	52% (95% CI: 48%–56%)	35% (95% CI: 31%–39%)
Time efficiency (avg. time saved)	33 min (SD = 5.7)	33 min (SD = 5.2)	31 min (SD = 6.1)
Editor satisfaction	7.2/10	7.8/10	6.5/10
AI suggestions deemed suitable	37%	41%	33%
Fictional reviewers suggested	10%	8%	12%
Effect size (Cohen's *d*)	0.62	0.68	0.35

1. Select a recent manuscript submitted to their journal that required peer review.
2. Use their traditional methods to identify potential reviewers for the manuscript, documenting their selections and the time taken.
3. Utilize the provided AI prompt with the same manuscript information to generate reviewer recommendations.
4. Compare the AI-generated list with their own selections, evaluating the relevance and suitability of the AI recommendations.

Measures and instruments

Data were collected through a structured online questionnaire that captured the following key metrics:

1. Selection Accuracy: The degree of overlap between AI recommendations and the editor's original selections.
2. Selection Quality: An assessment of whether the AI recommendations included reviewers that the editor had not considered but found suitable.
3. Time Efficiency: The time taken to find suitable reviewers using traditional methods compared with the AI system.
4. Editor Satisfaction: A rating of the AI recommendations' usefulness on a scale of 1–10.
5. Fictional Reviewers: Whether the AI system suggested any non-existent reviewers.

These metrics were chosen to provide a comprehensive evaluation of the AI system's performance, balancing quantitative measures of accuracy and efficiency with qualitative assessments of reviewer suitability and editor satisfaction. The inclusion of the 'Fictional Reviewers' metric was designed to assess the reliability and potential limitations of the AI system. Additionally, open-ended questions were included to gather qualitative feedback on the strengths and limitations of the AI system, including its potential to supplement existing selection methods. The complete questionnaire used to collect data from editors is available in Appendix B.

In this study, we defined selection accuracy as the agreement between AI-suggested reviewers and those considered viable options by editors. Specifically, this metric represents the number of reviewers (minimum 1 suitable reviewer) suggested by the AI that also appeared in the editors' list of reasonable choices from the pool of all qualified potential reviewers. While this approach provides a quantitative measure of AI performance, it is crucial to acknowledge its limitations. It does not account for the potential variability among editors in reviewer selection, nor does it consider the full range of suitable reviewers that may exist beyond the editors' immediate considerations. This highlights the complexity of establishing a true 'gold standard' in reviewer selection and suggests an important area for future research, potentially involving a more comprehensive comparison of AI suggestions against a broader consensus of expert opinions.

As this study explores an emerging area of research, the questionnaire used has not undergone formal validation, which is common in initial investigations of new phenomena (Edmondson & McManus, 2007). While this approach allows for flexibility in exploring new concepts, it may affect the reliability and validity of the data collected. Future studies should focus on developing and validating standardized measures for assessing AI-assisted peer-review processes.

Data analysis methods

Quantitative analysis:

- Paired *t*-tests were conducted to compare the time efficiency between the AI and traditional methods.
- Descriptive statistics were calculated for the selection accuracy, quality, and editor satisfaction ratings.

Qualitative analysis: The open-ended responses were subjected to thematic analysis using the constant comparative method (Glaser & Strauss, 1967). Two independent coders identified recurring themes, with discrepancies resolved through discussion to ensure inter-rater reliability. To account for potential discipline-specific variations, subgroup analyses were performed to compare AI performance across different academic fields. This approach allowed for a nuanced understanding of the AI system's efficacy in various scholarly contexts.

RESULTS

Quantitative findings

1. Selection accuracy: The AI-generated recommendations demonstrated a measurable level of overlap with editors' traditional selections. On average, in 42% of cases, the AI suggestions included at least one reviewer (and often 2–3 reviewers) that matched those independently chosen by editors. This overlap indicates that the AI system was able to identify reviewers that editors considered appropriate, validating its potential utility in the selection process (Table 2 summarizes the key performance metrics of the AI-assisted reviewer selection process). The accuracy varied across disciplines, with higher rates of agreement in fields like Chemistry (58%) and lower in areas like Public Health (31%). It is important to emphasize that this overlap represents instances where the AI's suggestions included one or more names that the editors had already considered as viable options, demonstrating the AI's ability to identify relevant experts across various fields.
2. Selection quality: Editors reported that 37% of AI-suggested reviewers who were not in their original selection were considered suitable upon review. However, 25% of AI suggestions were deemed too senior in their field, potentially making them less likely to accept review invitations.
3. Time efficiency: The AI system demonstrated significant time savings. On average, editors spent 45 min selecting reviewers using traditional methods, compared to 12 min using the AI system.[1] This represents a 73% reduction in time spent on reviewer selection.
4. Editor satisfaction: Editor satisfaction with the AI system varied, with an average rating of 7.2 out of 10. Satisfaction was higher among editors from STEM (Science, Technology, Engineering, and Mathematics) fields (average 7.8) compared to those from social sciences and humanities (average 6.5). Further analysis revealed significant disciplinary variations in AI performance. In STEM fields, the average overlap between AI-suggested reviewers and editors' selections was 52% (95% CI: 48%–56%), compared with 35% (95% CI: 31%–39%) in social sciences and humanities. The effect size (Cohen's *d*) for this disciplinary difference was 0.68, indicating a moderate to large effect. Time efficiency gains were consistent across disciplines, with mean time savings of 33 min ($SD = 5.2$) in STEM fields and 31 min ($SD = 6.1$) in social sciences and humanities.

Qualitative insights

The thematic analysis of editor feedback revealed several key insights into the AI-assisted reviewer selection process:

1. Identification of lesser-known experts: Editors expressed appreciation for the system's capacity to identify lesser-known experts (for them), potentially broadening the pool of reviewers. For example, one editor in Chemistry noted, 'The AI suggested two early-career researchers I had not considered, both with highly relevant recent publications'.
2. Limitations in interdisciplinary understanding: Concerns were raised about the AI's understanding of interdisciplinary work, highlighting a limitation in its ability to navigate complex, cross-disciplinary research areas. An editor in Sociology commented, 'The AI seemed to miss important methodological distinctions within our field, suggesting reviewers with quantitative expertise for a qualitative study'.
3. Mixed opinions on suitability assessment: Editors held mixed opinions regarding the system's capability to assess reviewer suitability beyond subject expertise, suggesting that the AI may not fully capture factors such as methodological alignment and theoretical perspective.
4. Complementary tool perception: Editors generally perceived the AI system as a valuable complement to traditional methods rather than a replacement. Many reported that AI suggestions helped expand their pool of potential reviewers and occasionally highlighted overlooked experts in niche sub-fields.
5. Fictional reviewer suggestions: While infrequent, instances of the AI system suggesting fictional reviewers were noted. This occurred in approximately 10% of cases, primarily in rapidly evolving fields where the AI's training data may have been outdated.
6. Over-suggestion of senior researchers: A quarter of the editors noted that the AI occasionally suggested reviewers who were too senior or well-known in their field to be realistic choices for ad hoc reviews.
7. Disciplinary variations: In disciplines such as Law and Sociology, editors reported that the AI sometimes failed to account for important methodological or theoretical distinctions within the field. Conversely, editors in areas like Public Health found the AI suggestions particularly beneficial in identifying experts from adjacent disciplines they might not have considered otherwise.
8. Potential bias concerns: A small proportion of editors (10%) expressed concerns about potential bias in the AI's suggestions, noting a perceived skew towards researchers from well-known institutions in North America and Europe.

DISCUSSION

Interpretation of key findings

Our study reveals a nuanced picture of AI-assisted reviewer selection in academic publishing. The level of overlap (42%) between AI-suggested reviewers and editors' selections indicates that while AI can identify relevant reviewers, it does not entirely replicate human expertise. The higher accuracy in fields like Chemistry (58%) compared to areas like Public Health (31%) suggests that AI performance varies across disciplines, likely due to differences in the structure and clarity of research boundaries.

[1]It is important to say that this time does not include the direct appeal to reviewers (such as emails, phone calls, etc.) but only the time needed to locate possible reviewers.

While these findings are encouraging, it is crucial to interpret them cautiously, considering a key limitation in our methodology. The 42% overlap between AI and editor selections, representing an agreement on a minimum of one suitable external ad hoc reviewers from a larger pool, is notable but should be contextualized. We used editors' selections as the benchmark for AI performance, but we did not investigate the potential variability among editors in reviewer selection. This raises important questions about what constitutes a 'gold standard' in this context. Future research should examine the inter-editor agreement rate to provide a more comprehensive baseline for evaluating AI performance. This could involve comparing selections between different editors or conducting a larger-scale study to establish a more robust measure of human expert performance in reviewer selection.

Moreover, the variation in AI performance across disciplines (58% in Chemistry vs. 31% in Public Health) suggests that the effectiveness of AI in reviewer selection may depend on the nature and structure of the field. This variability warrants further investigation and may have implications for how AI tools are developed and implemented across different academic disciplines. Our study did not include a systematic verification of the current affiliations provided by the AI for suggested reviewers. This is a limitation of the current study and an important area for future research. Accurate affiliation information is crucial for editor decision-making and could impact the overall effectiveness of AI-assisted reviewer selection.

The contrast between 42% accuracy and high time efficiency (73% reduction) raises important questions about the trade-offs between speed and precision in reviewer selection. While the time-saving benefit is clear, we must consider if this efficiency compromises selection quality. However, it is noteworthy that 37% of AI-suggested reviewers not initially considered by editors were deemed suitable. This suggests that AI is broadening the reviewer pool beyond editors' immediate networks—a valuable contribution to diversity in peer review.

The discrepancy in satisfaction levels between STEM (7.8/10) and social sciences/humanities editors (6.5/10) may be attributed to the different nature of research in these fields. STEM disciplines often have more clearly defined research boundaries and methodologies, which might align better with the capabilities of current AI systems. In contrast, the more interpretative and interdisciplinary nature of social sciences and humanities research may pose more significant challenges for AI systems.

Implications for academic publishing

The results suggest that AI-assisted reviewer selection has the potential to significantly streamline the editorial process, particularly in terms of time efficiency. This could help address the increasing burden on editors, as Kovanis et al. (2016) noted. However, the varying performance across disciplines implies that a one-size-fits-all approach to AI implementation may not be appropriate. The AI's potential to increase opportunities for early-career researchers in peer review seems at odds with its tendency to suggest overly senior reviewers. This apparent contradiction likely stems from the system's current limitations in balancing expertise with the career stage.

While our results demonstrate the potential benefits of AI-assisted reviewer selection, it is essential to address potential counterarguments. Critics may argue that AI systems could perpetuate existing academic biases or reduce the human element in the peer-review process. However, our findings suggest that AI can actually broaden the pool of potential reviewers and identify experts that editors might have overlooked. Nevertheless, we acknowledge the need for ongoing monitoring and refinement of AI systems to ensure they promote diversity and fairness in the peer-review process.

Strengths and limitations of AI in reviewer selection

The implementation of AI in reviewer selection demonstrates several strengths and limitations. First, in terms of strengths, AI significantly enhances efficiency, resulting in substantial time savings during the reviewer selection process. Second, it expands the reviewer pool by identifying relevant experts beyond editors' immediate networks. Third, AI ensures consistency by applying uniform criteria across all manuscripts.

However, the system also exhibits certain limitations. First, it demonstrates less accuracy in interdisciplinary fields, highlighting disciplinary variations in performance. Second, AI struggles with contextual understanding, particularly in grasping nuanced methodological or theoretical distinctions. Third, there is a tendency for over-suggestion of senior researchers, with the AI occasionally proposing reviewers too prominent to be realistic choices. The suggestion of fictional reviewers in 10% of cases underscores the necessity for human oversight. This occurrence raises important questions about the AI system's reliability and highlights a significant area for improvement. The fact that 1 in 10 suggested reviewers were fictional emphasizes the current limitations of the AI system and the crucial role of human verification in the reviewer selection process.

It is important to note that this study used a generally available AI model that was not fine-tuned for specific disciplines. This limitation may partly explain the performance variations observed between STEM and Social Sciences and humanities fields. Future research could explore the impact of discipline-specific fine-tuning on AI performance in reviewer selection. Such fine-tuning could potentially address issues like the over-selection of senior researchers and under-representation of certain groups.

Ethical considerations and potential biases

The integration of AI in reviewer selection raises several ethical considerations. First, the issue of algorithmic bias emerges, with a perceived skew towards researchers from well-known North American and European institutions, suggesting potential geographical and institutional biases in the AI's training data. Second, while AI has the potential to increase diversity in reviewer selection, it may inadvertently perpetuate existing biases if not carefully designed and monitored.

Third, the 'black box' nature of AI decision-making processes could reduce transparency in the reviewer selection process, a concern echoed by Horbach and Halffman (2020). Fourth, using AI to analyse researcher profiles raises essential questions about data privacy and consent, as highlighted by Teixeira da Silva et al. (2019). Last, the study emphasizes the need for human oversight in AI-driven reviewer selection, especially in interdisciplinary contexts, aligning with existing ethical concerns in the literature (Rodríguez-Bravo et al., 2017).

LIMITATIONS AND FUTURE RESEARCH

Study limitations

This study offers valuable insights into AI-assisted reviewer selection, but several limitations must be acknowledged. First, the sample size and diversity are limited. While the study included 20 editors from various disciplines, this may not fully represent the global academic publishing landscape, constraining the generalizability of the findings. Second, the evaluation was based on a single instance per editor, not capturing the full spectrum of reviewer selection challenges. Third, the study provides only a snapshot of AI performance, not assessing long-term impacts on the peer-review process or publication quality. Fourth, potential bias in the AI's training data is a concern, potentially favouring certain demographic groups or institutions. This potential bias is particularly concerning given the global nature of academic publishing. The AI system's training data, while extensive, may not fully represent the diversity of global scholarship, especially from regions with less digitized or accessible research outputs. This could lead to systemic underrepresentation of scholars from certain geographical areas or institutions in reviewer recommendations.

Fifth, the AI system's ability to understand nuanced disciplinary contexts, particularly in interdisciplinary fields, was not fully explored. Lastly, time efficiency data relied on editors' self-reporting, which may be subject to recall bias or imprecision.

Suggestions for future investigations

To address these limitations and advance our understanding of AI in academic publishing, we propose several avenues for future research. First, longitudinal studies will be conducted to assess the long-term impact of AI-assisted reviewer selection on peer-review quality, publication outcomes, and citation rates. Second, expand future studies to include more extensive, more diverse samples encompassing a broader range of countries, institutions, and academic disciplines. Third, multi-manuscript evaluations should be implemented to capture a broader range of reviewer selection scenarios. Fourth, conduct focused studies on AI performance in interdisciplinary fields to improve cross-disciplinary reviewer suggestions. Last, work towards establishing ethical guidelines and best practices for the use of AI in peer-review processes to ensure responsible and fair implementation.

A valuable metric for future studies would be to compare the review invitation acceptance rates between AI-suggested reviewers and those selected through traditional methods. This comparison could provide insights into the practical efficiency gains of AI-assisted reviewer selection and its potential to alleviate the burden on frequently invited, high-performing reviewers.

Long-term implications for academic publishing

The integration of AI into reviewer selection processes has the potential to significantly reshape academic publishing. First, AI could enhance efficiency and scale, enabling journals to handle increasing submission volumes more efficiently and potentially reducing publication delays. Second, AI might contribute to more diverse perspectives in the peer-review process by identifying a broader pool of potential reviewers. Third, the role of editors may evolve, shifting towards more strategic oversight and quality control of AI-assisted processes. Fourth, there may be a tension between the standardization AI brings and the need for discipline-specific customization in reviewer selection.

Fifth, AI systems could increase opportunities for early-career researchers to participate in peer review if designed to consider a broader range of expertise levels. Sixth, AI could help address geographical imbalances in reviewer selection, though this depends on the diversity of its training data. Seventh, the definition and identification of 'expert reviewers' may evolve with AI's ability to analyse publication patterns and emerging research trends. Last, the long-term use of AI in academic publishing will necessitate ongoing discussions about privacy, consent, and the appropriate balance between human and machine decision-making in scholarly communication. Developing standardized evaluation frameworks for assessing AI performance and impact will be crucial. Collaborative efforts between journal editors, AI researchers, and ethicists will be essential in developing these standards and ensuring responsible AI implementation in scholarly communication.

While AI-assisted reviewer selection shows promise, its successful integration into academic publishing will require careful consideration of these limitations and long-term implications. Future research should focus on refining AI systems to better understand disciplinary nuances, mitigate biases, and enhance the overall quality and efficiency of the peer-review process.

CONCLUSIONS

This study makes several significant contributions to the field of academic publishing and AI applications. First, it empirically demonstrates the efficiency gains and benefits of AI-assisted reviewer selection across disciplines. Second, it underscores the necessity of human oversight and tailored AI models for different fields. Third, it addresses key challenges and ethical concerns crucial for

AI's responsible integration into academic publishing. Our findings reveal a complex interplay between technological capabilities and the nuanced demands of scholarly peer review. The AI system achieved a 42% overlap with editors' choices, illustrating both its promise and inherent limitation. Notably, the system's performance varied across disciplines, performing more robustly in STEM fields compared to the social sciences and humanities. This variation underscores the need for discipline-specific considerations in developing and applying AI tools in academic publishing.

Perhaps the most notable finding is the 73% reduction in reviewer selection time, a substantial efficiency gain with far-reaching implications for reducing the burden on editors and accelerating the peer-review process. Moreover, the AI system's ability to suggest suitable reviewers not initially considered by editors in 37% of cases points to its potential for broadening the pool of peer reviewers, which could contribute to greater diversity and fresh perspectives in the review process.

However, these promising results must be tempered with an acknowledgment of the system's limitations. The occasional suggestion of overly prominent or, in rare cases, fictional reviewers highlight the ongoing need for human oversight and the importance of editors' expertise in the reviewer selection process.

The practical implications of these findings for journal editors are significant. AI-assisted reviewer selection offers a powerful means to enhance efficiency in the editorial workflow, potentially allowing editors to redirect their time and expertise to other critical aspects of the publication process. However, the implementation of such systems should be approached with caution, particularly in interdisciplinary fields where the AI's performance is less robust. Editors should maintain a critical eye on AI recommendations, leveraging the technology's ability to broaden the reviewer pool while exercising their judgement to ensure the appropriateness and diversity of reviewer selections.

Looking forward, the integration of AI in peer-review processes should be guided by several key principles:

1. Phased implementation: A gradual approach, allowing for continuous refinement and adaptation, is advisable.
2. Development of AI models tailored to different fields of study could help address the varying performance across disciplines.
3. Transparency: Clear communication about the use of AI, both to authors and reviewers, will be crucial in maintaining trust in the peer-review process.
4. Bias monitoring: Regular checks for potential biases, whether geographic, institutional, or demographic, are essential to ensure that AI systems enhance rather than hinder diversity in academic publishing.
5. Ethical guidelines: The establishment of clear ethical guidelines addressing data privacy, consent, and algorithmic transparency will be paramount.
6. Editor training: Comprehensive training for editors on effectively using and interpreting AI recommendations will be crucial for successful integration.
7. System integration: AI tools should be seamlessly incorporated into existing editorial management systems to maximize adoption and efficiency.

AI-assisted reviewer selection heralds a new era in academic publishing, promising to revolutionize peer review with remarkable efficiency and inclusivity. Our study envisions a future where editors save time, uncover hidden experts, and make the peer-review process more dynamic and responsive, addressing the challenges of rapid research growth and reviewer fatigue Embracing this AI-enhanced future, we stand on the brink of a transformative era in scholarly communication. The synergy of human expertise and AI will propel academic publishing to new heights of efficiency, fairness, and innovation. Our findings chart a path forward, empowering the academic community to create an ethical, inclusive, and effective peer-review system for the 21st century and beyond. This study is a pivotal step towards a future where technology and human insight combine to advance knowledge more rapidly and equitably.

ACKNOWLEDGMENTS

I wish to express my deep gratitude to every editor who participated in this study. Your time and expertise were essential in understanding the integration of artificial intelligence in the reviewer selection process.

CONFLICT OF INTEREST STATEMENT

The author declares no conflict of interest.

REFERENCES

Arns, M. (2014). Open access is tiring out peer reviewers. *Nature*, *515*(7528), 467. https://doi.org/10.1038/515467a

Csiszar, A. (2016). Peer review: Troubled from the start. *Nature*, *532*(7599), 306–308. https://doi.org/10.1038/532306a

Edmondson, A. C., & McManus, S. E. (2007). Methodological fit in management field research. *Academy of Management Review*, *32*(4), 1246–1264. https://doi.org/10.5465/amr.2007.26586086

Gasparyan, A. Y., Gerasimov, A. N., Voronov, A. A., & Kitas, G. D. (2015). Rewarding peer reviewers: Maintaining the integrity of science communication. *Journal of Korean Medical Science*, *30*(4), 360–364. https://doi.org/10.3346/jkms.2015.30.4.360

Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Aldine Publishing Company.

Godlee, F., & Jefferson, T. (1999). Peer review in health sciences.

Grimaldo, F., Marušić, A., & Squazzoni, F. (2018). Fragments of peer review: A quantitative literature analysis (1969-2015). *PLoS One*, *13*(2), e0193148. https://doi.org/10.1371/journal.pone.0193148

Habibzadeh, F. (2023). The future of scientific journals: The rise of UniAI. *Learned Publishing*, *36*(2), 326–330. https://doi.org/10.1002/leap.1514

Heaven, D. (2018). AI peer reviewers unleashed to ease the publishing grind. *Nature*, *563*(7733), 609–610. https://doi.org/10.1038/d41586-018-07245-9

Horbach, S. P. J. M., & Halffman, W. (2018). The changing forms and expectations of peer review. *Research Integrity and Peer Review*, *3*(1), 8. https://doi.org/10.1186/s41073-018-0051-5

Horbach, S. P. J. M., & Halffman, W. (2020). Innovating editorial practices: Academic publishers at work. *Research Integrity and Peer Review*, *5*(1), 11. https://doi.org/10.1186/s41073-020-00097-w

Kovanis, M., Porcher, R., Ravaud, P., & Trinquart, L. (2016). The global burden of journal peer review in the biomedical literature: Strong imbalance in the collective enterprise. *PLoS One*, *11*(11), e0166387. https://doi.org/10.1371/journal.pone.0166387

Kovanis, M., Trinquart, L., Ravaud, P., & Porcher, R. (2017). Evaluating alternative systems of peer review: A large-scale agent-based modeling approach to scientific publication. *Scientometrics*, *113*(1), 651–671. https://doi.org/10.1007/s11192-017-2375-1

Mrowinski, M. J., Fronczak, P., Fronczak, A., Ausloos, M., & Nedic, O. (2017). Artificial intelligence in peer review: How can evolutionary computation support journal editors? *PLoS One*, *12*(9), e0184711. https://doi.org/10.1371/journal.pone.0184711

Mulligan, A., Hall, L., & Raphael, E. (2013). Peer review in a changing world: An international study measuring the attitudes of researchers. *Journal of the American Society for Information Science and Technology*, *64*(1), 132–161. https://doi.org/10.1002/asi.22798

Price, S., & Flach, P. A. (2017). Computational support for academic peer review: A perspective from artificial intelligence. *Communications of the ACM*, *60*(3), 70–79. https://doi.org/10.1145/2979672

Publons. (2018). Global State of Peer Review. https://publons.com/community/gspr

Rivera, H. (2018). Fake peer review and inappropriate authorship are real evils. *Journal of Korean Medical Science*, *34*(2), 1–7. https://doi.org/10.3346/jkms.2019.34.e6

Rodríguez-Bravo, B., Nicholas, D., Herman, E., Boukacem-Zeghmouri, C., Watkinson, A., Xu, J., & Świgoń, M. (2017). Peer review: The experience and views of early career researchers. *Learned Publishing*, *30*(4), 269–277. https://doi.org/10.1002/leap.1111

Severin, A., & Chataway, J. (2021). Purposes of peer review: A qualitative study of stakeholder expectations and perceptions. *Learned Publishing*, *34*(2), 144–155. https://doi.org/10.1002/leap.1336

Severin, A., Strinzel, M., Egger, M., Domingo, M., & Barros, T. (2020). Who reviews for predatory journals? A study on reviewer characteristics. *BioRxiv*, *3*, 2–7.

Teixeira da Silva, J. A., Dobránszki, J., Tsigaris, P., & Al-Khatib, A. (2019). Predatory and exploitative behavior in academic publishing: An assessment. *The Journal of Academic Librarianship*, *45*(6), 102071. https://doi.org/10.1016/j.acalib.2019.102071

Tennant, J. P., Crane, H., Crick, T., Davila, J., Enkhbayar, A., Havemann, J., Kramer, B., Martin, R., Masuzzo, P., Nobes, A., Rice, C., Rivera-López, B., Ross-Hellauer, T., Sattler, S., Thacker, P. D., & Vanholsbeeck, M. (2019). Ten hot topics around scholarly publishing. *Publications*, *7*(2), 34. https://doi.org/10.3390/publications7020034

Tennant, J. P., Dugan, J. M., Graziotin, D., Jacques, D. C., Waldner, F., Mietchen, D., Elkhatib, Y., Collister, B., Pikas, C. K., Crick, T., Masuzzo, P., Caravaggi, A., Berg, D. R., Niemeyer, K. E., Ross-Hellauer, T., Mannheimer, S., Rigling, L., Katz, D. S., Greshake Tzovaras, B., … Colomb, J. (2017). A multi-disciplinary perspective on emergent and future innovations in peer review. *F1000Research*, *6*, 1151. https://doi.org/10.12688/f1000research.12037.3

APPENDIX

A.1. AI PROMPT USED IN THE STUDY

The following prompt was provided to the AI system (GPT-4) for each manuscript:

'You are an expert in academic publishing across various disciplines. I will provide you with the title, abstract, and keywords of an academic manuscript. Your task is to suggest 5–7 potential reviewers for this manuscript. For each suggested reviewer, provide:

1. Full name
2. Current affiliation
3. A brief explanation of why they would be suitable (2–3 sentences)
4. 1–2 relevant publications by this reviewer

Please ensure that the suggested reviewers are diverse regarding geography, career stage, and sub-specialties within the field. Do not suggest reviewers who are likely to have a conflict of interest with the manuscript's authors.

Here is the manuscript information: Title: [Insert manuscript title] Abstract: [Insert manuscript abstract] Keywords: [Insert keywords].

Please provide your reviewer suggestions based on this information'.

APPENDIX

B.1. QUESTIONNAIRE FOR EDITORS

1. How many of the AI-suggested reviewers match your original selections? (Numeric response)
2. On a scale of 1–10, how relevant were the AI-suggested reviewers to the manuscript topic? (1 = Not at all relevant, 10 = Extremely relevant)
3. How many of the AI-suggested reviewers that were not in your original selection would you consider suitable? (Numeric response)
4. Approximately how much time (in minutes) did you spend selecting reviewers using your traditional method? (Numeric response)
5. Approximately how much time (in minutes) did you spend reviewing the AI suggestions? (Numeric response)
6. On a scale of 1–10, how satisfied are you with the AI-suggested reviewers? (1 = Not at all satisfied, 10 = Extremely satisfied)
7. Did the AI system suggest any reviewers that do not exist or seem fictional? If yes, how many? (Yes/No, Numeric response if Yes)
8. What strengths did you observe in the AI-suggested reviewer list? (Open-ended response)
9. What limitations or weaknesses did you observe in the AI-suggested reviewer list? (Open-ended response)

10. How do you think AI-assisted reviewer selection could complement your current selection process? (Open-ended response)

APPENDIX

C.1. CHATGPT-4 (OPENAI)

ChatGPT-4 (OpenAI) Architecture: ChatGPT-4 is built on the generative pre-trained transformer (GPT) architecture, a sophisticated neural network model designed for NLP. While the exact number of parameters for GPT-4 has not been publicly disclosed, it is widely acknowledged that it is significantly larger than GPT-3.5, which had 175 billion parameters. This increase in scale enhances the model's ability to understand and generate complex language structures (OpenAI, 2024).

Training Process: GPT-4's training process involves two main stages: pre-training and fine-tuning. During pre-training, the model is exposed to a vast corpus of internet text data, which helps it learn the statistical patterns of language. This is followed by fine-tuning using more specific, curated datasets to improve its performance on targeted tasks. A key feature of GPT-4 is the use of Reinforcement Learning from Human Feedback (RLHF), where human evaluators provide feedback on the model's outputs, guiding it towards more accurate and contextually appropriate responses. This approach builds on the InstructGPT framework, enhancing GPT-4's ability to follow user instructions more effectively (OpenAI, 2024).

Commercial Availability: GPT-4 was introduced for commercial use in March 2024 as part of OpenAI's suite of AI tools, including its integration into the ChatGPT product available via subscription on the OpenAI platform (OpenAI, 2024). This commercial release made GPT-4 accessible to developers and businesses, offering enhanced capabilities for a wide range of applications such as customer service, content generation, and complex problem-solving.

OpenAI. (2024). Introducing GPT-4. Retrieved from https://openai.com/research/gpt-4.

OpenAI. (2024). API Documentation. Retrieved from https://platform.openai.com/docs.