Journal of the Association for Information Systems

Volume 25
Issue 1 *Special Issue: The Future Impact of AI on Academic Journals and the Editorial Process (pp. 37-181)*

Article 8

2024

The Other Reviewer: RoboReviewer

Ron Weber
Monash University / The University of Queensland, Ron.Weber@monash.edu

Recommended Citation

Weber, Ron (2024) "The Other Reviewer: RoboReviewer," *Journal of the Association for Information Systems*, 25(1), 85-97.
DOI: 10.17705/1jais.00866
Available at: https://aisel.aisnet.org/jais/vol25/iss1/8

Journal of the Association for Information Systems (2024) **25**(1), 85-97
doi: 10.17705/1jais.00866

EDITORIAL

ISSN 1536-9323

The Other Reviewer: RoboReviewer

Ron Weber[1]
[1]Faculty of Information Technology, Monash University /
School of Business, The University of Queensland, Australia, ron.weber@monash.edu

Abstract

The peer review process is a mainstay for informing publication decisions at many journals and conferences. It has several strengths that are well-accepted, such as providing a signal about the quality of published papers. Nonetheless, it has several limitations that have been documented extensively, such as reviewer biases affecting paper appraisals. To date, attempts to mitigate these limitations have had limited success. Accordingly, I consider how developments in artificial intelligence technologies—in particular, pretrained large language models with downstream fine-tuning—might be used to automate peer reviews. I discuss several challenges that are likely to arise if these systems are built and deployed and some ways to address these challenges. If the systems are deemed successful, I describe some characteristics of a highly competitive, lucrative marketplace for these systems that is likely to emerge. I discuss some ramifications of such a marketplace for authors, reviewers, editors, conference chairs, conference program committees, publishers, and the peer review process.

Keywords: Peer Review, Artificial Intelligence, Large Language Models, Pretraining, Fine-Tuning, Robot Reviewer Marketplace, Peer Review Ramifications

Dov Te'eni was the accepting senior editor. This paper was submitted on June 17, 2023 and underwent one revision. It is part of the Special Issue on The Future Impact of AI on Academic Journals and the Editorial Process.

1 Introduction

The peer review process, which is now used by many journals and conferences to inform their publication decisions, relies on experts on the topic of a submitted paper providing advice on its quality.[1] The process has several strengths. First, it helps to ensure only high-quality papers are published. Second, it assists researchers with limited time to decide which papers to read. Third, review feedback helps authors improve the quality of their papers and research work. Fourth, reviewers learn from the review work they undertake.

Nonetheless, the peer review process has been the focus of many criticisms (Table 1). Some underpin concerns about false negative decisions (wrongly rejecting papers) and false positive decisions (wrongly publishing papers) and the consequent impact on the quality of knowledge produced. Others underpin concerns about increasing overheads imposed on individual scholars and the scientific process overall.

[1] Introduction of the peer review process is often attributed to the Royal Society of Edinburgh and the Royal Society of London (e.g., Spier, 2002a, p. 357). Apparently, they commenced using the process in the mid-eighteenth century to evaluate the quality of manuscripts. However, Baldwin (2020, p. 1) argues research by historians shows the origins, forms, and evolution of the peer review process are nuanced and the term "peer review" did not appear until the latter part of the twentieth century.

Table 1. Example Criticisms of the Peer Review Process

Criticisms of the peer review process	Example references
Reviewer biases associated with the characteristics of a paper's author such as their likely identity, sex, institution, country, primary language	Jayasinghe et al., 2003; Lee et al.,, 2013; Resnik & Elmore 2016; Smith et al., 2023
Reviewers failing to identify research misconduct and significant errors in papers	Fanga et al., 2012; Mulligan et al., 2013; Nicholas et al., 2015; Schroter et al., 2008
Reviewers and editors preserving the status quo in a field	Lawrence 2003; Resch et al., 2000; Spier, 2002b
Reviewers recommending publication of positive findings but not negative findings	Resnik & Elmore, 2016; Smith, 2006
Reviewers failing to provide timely, constructive, high-quality reviews	Huisman & Smits, 2017; Smith, 2006
Reviewers lacking competence	Resnik et al., 2008
Editors giving only cursory justifications for their rejection decisions and abrogating their responsibilities to reviewers	Lawrence, 2003; Nicholas et al., 2015; Straub, 2008
Low interrater agreement among reviewers	Jackson et al., 2011; Lee, 2012
Reviewers stymieing progress by colleagues whom they deem to be competitors	Lawrence, 2003
Reviewers and editors failing to identify the importance and likely impact of papers	Bornmann & Daniel, 2009; Jackson et al., 2011; Starbuck, 2005; Straub, 2008
Ethical violations, such as reviewers appropriating authors' work and failing to protect the confidentiality of submissions	Resnik, 2011; Resnik et al., 2008; Smith, 2006
Difficulties finding reviewers because of the substantial growth in the number of scientific papers, journals, and conferences[a]	Fire & Guestrin, 2019; Johnson et al., 2018
Dealing with the "tragedy of the commons"—overconsumption by some scholars of peer reviews relative to their contribution to peer review work	Kautz, 2018; Kovanis et al., 2016; Sipior, 2018; Stafford, 2018
High cost of peer review system[b]	Aczel et al., 2021; LeBlanc et al., 2023

Note: [a]Johnson et al. (2018, pp. 25-28) estimate (a) about 33,100 peer reviewed English-language journals and 9,400 non-English-language journals publish over three million articles per year, and (b) the annual growth rate is 4% in articles and 5-6% in journals. [b]LeBlanc et al. (2023) estimate the cost of a peer review in 2020 was about 1,272 USD per person and the global cost was 6 billion USD.

Notwithstanding the problematic nature of the peer review process, research indicates overall satisfaction with it remains high (Johnson et al., 2018; Mulligan et al., 2013, p. 137; Nicholas et al., 2015). As a result, major work has been done to try to improve its quality and efficacy (Waltman et al., 2023). However, the evidence in support of this work improving peer review processes is mixed, weak, or nonexistent (e.g., Bravo et al., 2019; Jefferson et al., 2002; Smith, 2006; Spezi et al., 2018).

Recently, artificial intelligence (AI) systems have been developed to achieve better peer review process outcomes. For instance, AI systems have been built to assist (1) conference program chairs in assembling a high-quality program committee (e.g., Price & Flach, 2017), (2) journal editors and conference program chairs to select high-quality reviewers (e.g., Rahut et al., 2022), (3) editors to approach reviewers in an optimal sequence (Mrowinski et al., 2017), and (4) reviewers to obtain a summary of the key points in a paper and references to related research (e.g., Golan et al., 2023). Similarly, to facilitate the peer review process, AI systems have been developed to assist researchers to (1) improve the structure, grammar, style, and readability of their papers (e.g., Golan et al., 2023), (2) ground their research in higher-quality literature reviews (e.g., Wagner et al., 2022), and (3) choose the best publication outlet for their paper (e.g., Razack et al., 2021).[2]

In this opinion piece, I seek to contribute to this growing body of literature. My focus is the possibility of developing and using AI systems to undertake probably the most difficult task in the peer review process—

[2] For some time, AI tools have also been used extensively for plagiarism detection and management of manuscript submissions, review processes, editorial workflows, and publication processes (e.g., Razack et al., 2021).

namely, obtaining high-quality, timely peer reviews of papers. To the best of my knowledge, only cursory consideration has been given so far to the idea that an AI system—a RoboReviewer—could be developed that would undertake high-quality reviews of papers (e.g., Susarla et al., 2023; van Dis et al., 2023). Indeed, some researchers have dismissed the idea outright because they doubt peer review processes can ever be automated (e.g., Aczel et al., 2021; Schulz et al., 2022).[3] Moreover, some editors argue that the use of AI undermines the quality of peer review processes because critical activities are best done via human interactions (Gendron et al., 2022).

Nonetheless, with significant, rapid, ongoing developments occurring in AI tools and techniques,[4] I am optimistic that in due course AI will make significant inroads into the peer review process. Moreover, even if the use of AI for peer review eventually proves to be intractable, it behooves us as researchers to consider the possibility carefully so that we are better prepared for challenges that might arise (e.g., ethical challenges).

My opinion piece proceeds as follows. First, I examine the required competencies of a high-quality peer reviewer. Second, I consider how a RoboReviewer that possesses these competencies might be built. Third, I examine some broad challenges that would arise when building and using a RoboReviewer. Fourth, I canvass the possibility of a RoboReviewer marketplace arising and some implications for researchers, reviewers, editors, and publishers of journals and conference proceedings. Fifth, I consider some ways in which the availability of high-quality RoboReviewers might change the peer review process. Finally, I present some brief reflections and conclusions.

2 Peer Reviewer Core Competencies

Table 2 is an adaptation and summary of Table 4 in Barroga (2020, p. 8) that lists the required competencies of high-quality peer reviewers, defined by the Council of Science Editors, World Association of Medical Editors, International Committee of Medical Journal Editors, and Committee on Publication Ethics. The competencies mirror those given in many other places (e.g., Köhler et al., 2020, pp. 8-17; Spyns & Vidal, 2015, pp. 11-17).

AI-based software already exists that possesses some of these “competencies,” at least in part. For example, writing-assistance software can be used to identify problems with a manuscript’s grammar and style, improve its clarity and conciseness, and ensure relevant literature is cited (Golan et al., 2023). Nonetheless, if RoboReviewers are to be deemed competent reviewers, the challenge is to determine whether they can acquire all the required competencies. To what extent, therefore, do any of these competencies require human intelligence? Or will some or all succumb eventually to increasingly intelligent machines?

Table 2. Peer Reviewer Core Competencies

Responsibility to:	**Core competency**
Authors	• Provide timely, written feedback on a paper’s strengths and weaknesses. • Provide constructive feedback to enable improvements to research that underpins paper and paper itself. • Indicate whether narrative in paper is appropriate, clear, concise, relevant, and compelling. • To the extent possible, avoid bias in feedback provided. • Maintain confidentiality of review process and author’s intellectual work.
Editors	• Disclose any conflict of interest. • Follow the editor’s instructions. • Provide high-quality reviews to assist editors to write a report and reach a disposition decision on a paper. • Note any ethical concerns and potential conflicts of interest that exist pertaining to the research reported in a paper. • Inform editor of review expertise and experience.
Readers	• Ensure a paper reports high-quality research. • Ensure a paper cites relevant research.

[3] Typical concerns are that AI will not be able to evaluate whether authors have (1) chosen a high-quality research question, (2) investigated it using appropriate research methods, (3) gathered appropriate data, and (4) reached conclusions that are supported by the data they have analyzed and interpreted. See, e.g., https://www.wired.com/2017/02/ai-can-solve-peer-review-ai-can-solve-anything/

[4] Many years ago, Moravec (1998) argued that the performance of AI machines improved “at the same pace” as improvements in hardware processing speed and memory capacity. These improvements have continued unabated.

3 How Might RoboReviewers be Built?

Two ways in which RoboReviewers might be built are to (1) emulate (mirror) human intelligence and (2) employ AI tools and techniques to implement useful applications (Shneiderman, 2020a, pp. 74-76). The former approach seeks to produce machines that have the same cognitive capabilities as humans—artificial *general* intelligence (AGI). They should be capable of achieving all the peer review competencies shown in Table 2. The latter approach has the more modest goal of producing machines that rely in part on knowledge of human intelligence but also employ tools and techniques that have proved useful in equipping machines to work on complex problems and tasks—artificial *narrow* intelligence (ANI). Most likely, they would be capable of achieving only some of the peer review competencies shown in Table 2 (either in whole or in part).

For three reasons, I doubt AGI will be achieved in the foreseeable future. First, the nature of human consciousness will need to be much better understood—what Chalmers (1996) calls the "hard problem." After hundreds of years of research in philosophy and neuroscience, the "hard problem" remains an enigma. Second, I am persuaded by the arguments of philosophers who characterize the human mind as an *emergent property* of the human body—a property that depends on properties of the human body's components (atoms, organelles, cells, tissues, organs, etc.) in often unknown, complex, seemingly impenetrable ways (e.g., Bunge, 1977; Mahner, 2015). Third, human consciousness, cognition, and intelligence also appear to be "embodied" in human sensorimotor capabilities—how humans perceive the world, sense its dynamics, and move around in it (Shapiro & Spaulding, 2021; Wilson, 2002). If these sensorimotor capabilities are not provided by human biological components, however, some base properties that underpin the human mind as an emergent property are likely to be missing (Mahner, 2015, p. 193).

Furthermore, even if AGI could be achieved, it is not always an appropriate goal when the limitations of human intelligence undermine task performance. For instance, Korteling et al. (2021, pp. 4-5) highlight task-performance problems that can arise because of (1) the small capacity of a human's short-term memory (Miller, 1956) and (2) well-known human cognitive biases, such as anchoring and adjustment and the recency effect (Tversky & Kahneman, 1974). During peer review processes, these limitations might detract from a human's ability to provide high-quality peer reviews (e.g., reviewers being less open to new ideas because they are influenced unduly by papers they have read recently).

For these reasons, I believe RoboReviewers must be built based on an ANI paradigm (at least in the short and medium term). Currently, the AI technology that appears most appropriate is a pretrained large language model (LLM) with downstream fine-tuning of the model (e.g., Qiu et al., 2020; Wang et al., 2022). For instance, to build a RoboReviewer for the *Journal of the Association for Information Systems* (JAIS), a pretrained LLM could be fine-tuned using the resources of the ScholarOne™ platform that JAIS employs to manage manuscripts.[5] ScholarOne™ now has an extensive database of manuscripts submitted to JAIS, their histories (e.g., eventual disposition, number of revisions), reviewer reports, associate editor and senior editor reports, and decision letters. This database could be used to retrain a pretrained LLM so that it is better suited for JAIS review purposes. Moreover, by incorporating a "continual-learning" capability into the fine-tuned LLM, review materials associated with new submissions could be used to further train the LLM.

4 Some Challenges in Building and Using RoboReviewers

Even with increasingly powerful AI tools and techniques, building and using a RoborReviewer successfully will not be straightforward. Below I reflect briefly on five challenges that will arise.

4.1 How Best to Use Human Reviewers in Conjunction with RoboReviewers?

For some time, I doubt a RoboReviewer will be able to act autonomously in a review process. When a new submission arrives, the RoboReviewer will first have to read it. What then occurs, however, is unclear. The following is one scenario that seeks to balance human control with automation to achieve trustworthy outcomes (Shneiderman, 2020b):

- After a preliminary examination of the submission, a senior editor (SE) assigns it to an associate editor (AE) who has expertise (1) in the topic of the submission, (2) priming and prompting the RoboReviewer, and (3) interpreting and evaluating the RoboReviewer's output.

[5] An LLM that has been pretrained using a smaller corpus of materials relevant to peer review tasks is likely to provide a better basis for a RoboReviewer than an LLM that has been pretrained on a large, diffuse corpus of materials.

- The AE then primes and prompts the fine-tuned LLM component of the RoboReviewer until they feel they have obtained the best feedback the RoboReviewer can provide.[6]
- The AE then interprets and evaluates the RoboReviewer's output to determine how to proceed. If they are concerned about the output, they may decide to engage a full panel of human reviewers. Alternatively, if they conclude the RoboReviewer has provided high-quality output, they might engage none, one, or a small number of human reviewers.
- If the AE decides to approach one or more human reviewers, they might use issues raised by the RoboReviewer to determine the kind of human reviewer expertise they need to evaluate the submission and how human reviewers should focus their review feedback.
- If human reviewer reports are obtained, the AE might decide to once again prime and prompt the RoboReviewer based on issues identified by the human reviewers.

If the peer review process is to be improved overall, the strengths of human reviewers and RoboReviewers somehow must be leveraged synergistically. On the one hand, humans are capable of empathy, intuition, perception, self-awareness, social awareness, and common-sense reasoning. These capabilities sometimes allow human reviewers who are experts in a field to have a deep understanding of underlying themes, contradictory theories and findings, inherent and problematic assumptions, and contributions that will have longevity rather than faddish appeal. On the other hand, while current AI systems lack these capabilities, a high-quality RoboReviewer used by an AE who has expertise in a paper's topic and priming and prompting expertise should be able to provide faster, more comprehensive reviews than humans and reviews that are less subject to the vagaries of human emotions and cognitive limitations.

4.2 Need for a Comprehensive, Representative Platform Database for Fine-Tuning Purposes

The quality of reviews provided by a RoboReviewer will depend in part on the quality of the platform database used to fine-tune it. The database needs to contain a comprehensive, representative set of varying-quality submissions and varying-quality review packages (review reports, editorial reports, and editorial correspondence).

For instance, if a journal publishes literature-review papers, the platform database must contain a reasonable number of paper submissions of varying quality that are representative of the different types of literature reviews that researchers might undertake (Paré et al., 2015). It must also contain review packages of varying quality relating to these papers. Otherwise, the RoboReviewer will be unable to create the predictive models it needs to generate high-quality reviews of new literature review submissions.

Similarly, if the publication policies of a journal primarily target quantitative research methods that have well-defined evaluation criteria (e.g., experiments, surveys, and econometrics analyses), a smaller platform database might be satisfactory for fine-tuning purposes. If a journal's focus is qualitative research, however, evaluation criteria are often less well-defined (Lee & Sarker, 2023). Thus, a larger platform database most likely will be needed to fine-tune a RoboReviewer so it can provide high-quality output.

A challenge that the developers of RoboReviewers face is that a single platform database might contain insufficient instances of the types of submissions and review packages needed to generate high-quality predictive models in an LLM. Somehow, several platform databases might have to be obtained and combined.

4.3 Dealing with Inherent Biases

While prima facie RoboReviewers should be less impacted by emotions, conflicts of interest, and biases, these human attributes may be embedded, sometimes subtly, in the pre- and retraining texts provided to RoboReviewers and thus affect the predictive models they use (Bender et al., 2021; Susarla et al., 2023; van Dis et al., 2023). For instance, platform databases might generate biases in an LLM against innovative, path-breaking papers because (1) they contain few instances of such papers and their associated review packages, and/or (2) the review packages that exist for such papers embed reviewer and editor biases toward preserving the status quo on a research topic.

One way the impact of some biases might be mitigated is to track the histories of submissions—for instance, a paper's impact indicators, such as the number of cites, downloads, and altmetrics (e.g., tweets, Facebook comments). For published papers, these histories should be relatively easy to construct. For rejected papers, tracing their history might be problematic, although some journals now ask authors to provide their paper's submission history. If a paper's history

[6] The ways in which LLM systems are primed and prompted are critical to the success of using these systems. "Prompt engineering" has now become a topic of significant interest—e.g., see https://www.promptingguide.ai. Indeed, Sejnowski (2023) conceives of a user's ability to prime and prompt an LLM as a "reverse Turing test." Via a user's priming and prompting actions, the AI system is finding out whether it is dealing with an intelligent human!

could be constructed, machine learning (ML) algorithms could then be used to determine whether certain characteristics of review packages are predictive of high-impact papers that were rejected (potentially false negative decisions) or low-impact papers that were accepted (potentially false positive decisions). Expert human reviewers could then examine the ML output to evaluate whether the platform database and the peer review processes used manifest biases.[7] If biases are apparent, feedback could be provided to the RoboReviewer to mitigate their effects.

4.4 Dealing with a Lack of Review Transparency

Many researchers are unlikely to trust RoboReviewers if they do not understand how they reach conclusions about the strengths or weaknesses of a paper. Insofar as the type of learning that underpins RoboReviewers is based on deeply layered neural networks, eliciting an explanation of their reasoning processes may be difficult (e.g., Hutson, 2018b; Korteling et al., 2021, p. 7; Wang et al., 2022).

Nonetheless, even human peer reviewers sometimes have difficulty articulating the rationales they have used to reach their conclusions about a paper's quality. Indeed, the more expert a person becomes at what they do, the greater the difficulty they have in explaining their actions—the "paradox of expertise" (Dror, 2011, p. 182). Accordingly, the quality of a RoboReviewer might be better judged based on the quality of its output rather than its ability to provide rationales for its conclusions.

Even if journal editors or conference program committee chairs believe the RoboReviewer they use produces high-quality output, however, researchers might still distrust it if its output is opaque (e.g., Checco et al., 2021, p. 9). Accordingly, researchers who receive unfavorable review decisions might be more likely to challenge a disposition decision (particularly if they can produce favorable output from another RoboReviewer). To reduce this possibility, the editors responsible for final paper disposition decisions must become adept at explaining their decisions in light of RoboReviewer output.

A significant problem that has occurred with some AI systems is the difficulty in reproducing their results (Hutson, 2018a). Prudent editors might therefore keep audit trails of the primes and prompts used with the RoboReviewer they employ to produce review reports. The value of these audit trails will be reduced as the RoboReviewer learns and evolves. Nonetheless, they might retain some value in the event that authors challenge an editor's decision and reproducing the output of a RoboReviewer becomes important to the editor-researcher exchanges that occur.

4.5 Being Cognizant of the Impacts on Science and Researcher

If RoboReviewers attain a prominent place in the peer review process, they will broadly affect the integrity of the scientific process because they influence important publication decisions.[8] Specifically, they will impact researchers' work lives and potentially their personal lives. For instance, if RoboReviewers do not provide high-quality output, researchers whose papers are wrongly rejected may become demoralized, have their careers affected negatively, and experience spillover effects in their personal lives.

Journal editors or conference program committee members who use a RoboReviewer must accept full responsibility for the feedback they provide to authors. If review errors, biases, or irregularities surface after they have provided feedback to authors, they cannot shift responsibility to the RoboReviewer without undermining their reputations and the reputation of the journal or conference for which they are responsible.

5 A RoboReviewer Marketplace?

If high-quality RoboReviewers can be built, I predict that a highly competitive, lucrative market for them (purchase, license, pay-per-use) will quickly emerge. Researchers have strong incentives to use one or more of them before submitting their papers to journals or conferences. Moreover, for their domain of expertise, researchers are likely to develop effective priming and prompting protocols.

Because substantial investments will be needed to develop high-quality RoboReviewers, large publishers are likely to develop them initially to support editors who work with their portfolio of journals and conference proceedings. They also have a comparative advantage in developing RoboReviewers because they have suites of platform databases they can use to pretrain and fine-tune a RoboReviewer's LLM.

Nonetheless, independent private providers have strong incentives to enter the market quickly. With late market entry, they may fear attracting researchers to their products will be difficult because switching costs are high. Furthermore, open source RoboReviewers may appear quickly because of concerns about (1) the opaqueness of and difficulties in reproducing results

[7] Storey et al. (2022, p. 29) warn that "uncomfortable organizational truths" may surface when these types of analyses are done.

[8] In some jurisdictions, the use of RoboReviewers may be subject to legislation because of concerns about harmful outcomes (Madiega, 2023).

with proprietary RoboReviewers and (2) reducing disadvantages experienced by researchers who lack the resources needed to purchase or access independent private-provider RoboReviewers (Spirling, 2023; van Dis et al., 2023, pp. 225-226).

Over time, the extent to which publishers retain a comparative advantage with their RoboReviewers because of their existing suite of platform databases is likely to be eroded. To reduce the power of publishers, researchers might be willing (or incentivized) to store their papers and review outcomes in public databases. Providers of independent and open source RoboReviewers could use these databases to train their systems.

Various open-publishing initiatives might also assist independent vendors of and sponsors of open source RoboReviewers to improve their products. For instance, open-access mega journals (e.g., the Public Library of Science's *PLOS ONE* and Nature's *Scientific Reports*) publish papers once they have passed a basic quality threshold. Post-publication peer reviews occur *directly* when interested researchers comment publicly on a paper and *indirectly* via measures such as a paper's number of citations, downloads, and altmetrics. Independent or open-source RoboReviewers could be trained using these resources.

The Manuscript Exchange Common Approach (MECA) (Lagace, 2019) may also assist development of independent and open source RoboReviewers. MECA provides standards that allow different manuscript management systems to exchange manuscripts and their related documents (e.g., reviews). Even though different journals and publishers compete with one another, some are adopting MECA to help reduce overall peer review costs. If providers of independent and open source RoboReviewers can gain access to the MECA ecosystem, they could use manuscript "packages" to pretrain or fine-tune their systems.

Some publishers may decide to make their RoboReviewers available to researchers who are considering whether to submit their papers to one of their journals or conferences (particularly if the publisher's comparative advantage with its RoboReviewers is being undermined). The publisher's goal might be to obtain fewer but higher-quality submissions to its journals and conferences. The nature and conduct of any review process that then occurs with submitted papers would need to be discerned.

6 Changed Paper Submission Dynamics

What outcomes might occur if researchers have access to high-quality RoboReviewers? Before researchers submit their papers to journals and conferences, presumably they will refine their papers based on not only "friendly reviews" from colleagues but also feedback from one or more RoboReviewers. As RoboReviewers evolve, they might also provide a quality rating for a researcher's paper and predictions of likely acceptance in different journals and conferences. Accordingly, one outcome might be that researchers reduce the number of papers they would otherwise submit to journals and conferences. Another might be that they better target those journals and conferences that are more likely to accept their papers. Both outcomes would reduce per capita peer review costs.

Presumably, prescreening activities at journals and conferences such as scans for plagiarism, doctored images, AI-generated content, and paper-mill output would also be rendered less effective (Hu, 2023; Tang, 2023). A RoboReviewer used by a researcher should have already detected these irregularities in their paper and possibly modified the paper to mask them.

The use of RoboReviewers might lead to more disputes about review outcomes. For instance, if a researcher's paper is rejected or they feel that egregious revisions are required as a condition for acceptance, they may use output from a RoboReviewer to challenge the review outcome.[9] The journal or conference affected will require protocols to handle such cases.

Researchers' use of RoboReviewers might also lead to a greater "homogenization" of research.[10] This outcome might occur because some researchers intentionally "game" RoboReviewers by choosing topics, theories, research methods, etc. that they believe are more likely to produce favorable review output by RoboReviewers. The use of RoboReviewers might also lead researchers unwittingly to shape their papers in particular ways. A consequence might be that, over time, papers manifest less creativity and innovation.

7 Conclusion

In this opinion piece, I have reflected on some possible futures where significant parts of the peer review process used in scholarly publishing are undertaken by AI systems. If this outcome were to arise, I predict the AI systems used will be applications of AI tools and techniques rather than emulations of human intelligence. I have discussed several challenges that are

[9] A problematic case arises if researchers have access to the journal's RoboReviewer or if the RoboReviewer used by authors has been trained with many of the same materials used to train the journal's RoboReviewer. How the RoboReviewer has been primed and prompted then becomes important to resolving disputes.

[10] I am indebted to a reviewer for pointing out this possible outcome.

likely to arise if these systems are deployed and some ways they might be addressed. I have also predicted that a highly competitive, lucrative marketplace will emerge for AI systems that are capable of undertaking key parts of the peer review process. The ramifications of this market will be significant for researchers, reviewers, editors, conference chairs, conference program committees, and publishers.

In my prognostications above, I have skirted a deep issue that lies at the heart of whether high-quality RoboReviewers can ever be developed—namely, whether LLMs (or some other technology) will ever achieve a form of understanding that allows RoboReviewers to attain high levels of peer review competencies. Current LLMs predict missing words from limited input based on deep neural networks that have billions and sometimes trillions of parameters (Bender et al., 2021, p. 611). These networks and parameters are the outcome of unsupervised machine learning that takes place using enormous amounts of natural-language text from many different sources.

Compelling arguments can be made that the high predictive ability of LLMs does not mean they understand natural language in the same ways as humans (e.g., Mitchell & Krakauer, 2023). Indeed, tests of LLMs often show their brittleness (e.g., Floridi & Chiriatti, 2020; Haman & Školník, 2023; Qureshi et al., 2023). For instance, they sometimes produce “hallucinations” in the form of erroneous and nonsensical output. Nonetheless, I am hopeful that the brittleness of RoboReviewers could be reduced in three ways: first, by pretraining its LLM using a contextualized database of a reasonable size (namely, one comprising multiple, relevant platform databases); second, by fine-tuning the LLM using the database of the platform it is intended to support; and third, by incorporating a continual-learning capability into the RoboReviewer.

Arguing that LLMs are simply highly complex predictive models is also an oversimplification of their nature. In this regard, Sejnowski (2023, pp. 317-319) argues that how LLMs *generalize* from what they learn potentially informs new theories of intelligence, consciousness, language acquisition and production, and how human brains work. Wei et al. (2022) also found evidence that LLMs acquire *emergent* properties when they are scaled up—these properties are not possessed by smaller-scale LLMs and cannot be predicted via extrapolations. Thus, the nature of the capabilities that will arise as LLM technologies improve is uncertain.

I see current developments with LLMs as a bellwether for AI applications that will make important inroads into the peer review process. Whatever the technology used, I doubt RoboReviewers will eliminate the need for human reviewers. Some type of “intelligent partnership” (van Dis et al., 2023, p. 226) will be needed. How this partnership should be enacted needs careful discernment.

Acknowledgments

I am indebted to Dov Te’eni, Roman Lukyanenko, and a reviewer for helpful comments on earlier versions of my opinion piece. Nonetheless, any errors or omissions in my opinion piece are my responsibility alone. I would also like to acknowledge the late Phillip Ein-Dor, the late Paul Gray, and Henry C. Lucas with whom I had the great pleasure and privilege of working during the establishment of the *Journal of the Association for Information Systems* and *Communications of the Association for Information Systems*. It was their hard work and inspiration that eventually led to the establishment of these two important journals.

References

Aczel, B., Szaszi, B., & Holcombe, A. O. (2021). A billion-dollar donation: Estimating the cost of researchers' time spent on peer review. *Research Integrity and Peer Review*, *6*(1), Article 14.

Baldwin, M. (2020). Peer review. In C. J. Phillips (Ed.), *Encyclopedia of the history of science* (pp. 1-15). PA: Carnegie Mellon University.

Barroga, E. (2020). Innovative strategies for peer review. *Journal of Korean Medical Science*, *35*(20), Article e138.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency* (610-623).

Bornmann, L., & Daniel, H.-D. (2009). Extent of type I and type II errors in editorial decisions: A case study on *Angewandte Chemie International Edition*. *Journal of Informetrics*, *3*(4), 348-352.

Bravo, G., Grimaldo, F., López-Iñesta, E., Mehmani, B., & Squazzoni, F. (2019). The effect of publishing peer review reports on referee behavior in five scholarly journals. *Nature Communications*, *10*, Article 322.

Bunge, M. A. (1977). Emergence and the mind. *Neuroscience*, *2*(4), 501-509.

Chalmers, D. J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford University Press. http://ebookcentral.proquest.com/lib/monash/detail.action?docID=272854

Checco, A., Bracciale, L., Loreti, P., Pinfield, S., & Bianchi, G. (2021). AI-assisted peer review. *Humanities and Social Sciences Communications*, *8*(1), Article 25.

Dror, I. E. (2011). The paradox of human expertise: Why experts get it wrong. In N. Kapur (Ed.), *The paradoxical brain* (pp. 177-188). Cambridge University Press.

Fanga, F. C., Steen, R. G., & Casadevalld, A. (2012). Misconduct accounts for the majority of retracted scientific publications. *PNAS*, *109*(42), 17028-17033.

Fire, M., & Guestrin, C. (2019). Over-optimization of academic publishing metrics: Observing Goodhart's Law in action. *GigaScience*, *8*(6), Article giz053.

Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, *30*(4), 681-694.

Gendron, Y., Andrew, J., & Cooper, C. (2022). The perils of artificial intelligence in academic publishing. *Critical Perspectives on Accounting*, *87*, Article 102411.

Golan, R., Reddy, R., Muthigi, A., & Ramasamy, R. (2023). Artificial intelligence in academic writing: A paradigm-shifting technological advance. *Nature Reviews Urology*. *20*, 327-328.

Haman, M., & Školník, M. (2023). Letter to the editor: Using ChatGPT to conduct a literature review. *Accountability in Research*. https://doi.org/10.1080/08989621.2023.2185514

Hu, G. (2023). Challenges for enforcing editorial policies on AI-generated papers. *Accountability in Research*, Advance online publication. https://doi.org/10.1080/08989621.2023.2184262

Huisman, J., & Smits, J. (2017). Duration and quality of the peer review process: The author's perspective. *Scientometrics*, *113*(1), 633-650.

Hutson, M. (2018a). Artificial intelligence faces reproducibility crisis. *Science*, *359*(6377), 725-726.

Hutson, M. (2018b). Has artificial intelligence become alchemy? *Science*, *360*(6388), Article 478.

Jackson, J. L., Srinivasan, M., Rea, J., Fletcher, K. E., & Kravitz, R. L. (2011). The validity of peer review in a general medicine journal. *PLOS ONE*, *6*(7), Article e22475.

Jayasinghe, U. W., Marsh, H. W., & Bond, N. (2003). A multilevel cross-classified modelling approach to peer review of grant proposals: The effects of assessor and researcher attributes on assessor ratings. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *166*(3), 279-300.

Jefferson, T., Alderson, P., Wager, E., & Davidoff, F. (2002). Effects of editorial peer review: A systematic review. *JAMA*, *287*(21), 2784-2786.

Johnson, R., Watkinson, A., & Mabe, M. (2018). *The STM Report: An overview of scientific and scholarly journal publishing* (5th ed.). STM: International Association of Scientific, Technical and Medical Publishers.

Kautz, K. (2018). Debate section editorial note: Reviews, reviewers, and reviewing: The "tragedy of the commons" in the scientific publication process. *Communications of the*

Association for Information Systems, *42*(1), 621-623.

Köhler, T., González-Morales, M. G., Banks, G. C., O'Boyle, E. H., Allen, J. A., Sinha, R., Woo, S. E., & Gulick, L. M. V. (2020). Supporting robust, rigorous, and reliable reviewing as the cornerstone of our profession: Introducing a competency framework for peer review. *Industrial and Organizational Psychology*, *13*(1), 1-27.

Korteling, J. E., van de Boer-Visschedijk, G. C., Blankendaal, R. A. M., Boonekamp, R. C., & Eikelboom, A. R. (2021). Human- versus artificial intelligence [Conceptual Analysis]. *Frontiers in Artificial Intelligence*, *4,* Article 622364.

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), Article e0166387.

Lagace, N. (2019). Manuscript exchange common approach: Supporting better publication flows. *Serials Review*, *45*(1-2), 88-89.

Lawrence, P. A. (2003). The politics of publication. *Nature*, *422*(6929), 259-261.

LeBlanc, A. G., Barnes, J. D., Saunders, T. J., Tremblay, M. S., & Chaput, J.-P. (2023). Scientific sinkhole: Estimating the cost of peer review based on survey data with snowball sampling. *Research Integrity and Peer Review*, *8*(1), Article 3.

Lee, A. S., & Sarker, S. (2023). A philosophical perspective on qualitative research in the age of digitalization. In B. Simeonova & R. D. Galliers (Eds.), *Cambridge handbook of qualitative digital research* (pp. 15-27). Cambridge University Press.

Lee, C. J. (2012). A Kuhnian critique of psychometric research on peer review. *Philosophy of Science*, *79*(5), 859-870.

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review [Article]. *Journal of the American Society for Information Science & Technology*, *64*(1), 2-17.

Madiega, T. (2023). *Artificial intelligence liability directive* (Briefing: EU Legislation in Practice). European Parliament. https://www.europarl.europa.eu/RegData/etudes/BRIE/2023/739342/EPRS_BRI(2023)739342_EN.pdf

Mahner, M. (2015). The philosophy of mind needs a better metaphysics. In S. M. Miller (Ed.), *The constitution of phenomenal consciousness: Toward a science and theory* (pp. 293-309). John Benjamins.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *The Psychological Review*, *63*(2), 81-97.

Mitchell, M., & Krakauer, D. (2023). The debate over understanding in AI's large language models. *PNAS*, *120*(13), Article e2215907120.

Moravec, H. (1998). When will computer hardware match the human brain? *Journal of Evolution and Technology*, *1*, 1-12.

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), Article 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.

Nicholas, D., Watkinson, A., Jamali, H. R., Herman, E., Tenopir, C., Volentine, R., Allard, S., & Levine, K. (2015). Peer review: Still king in the digital age. *Learned Publishing*, *28*(1), 15-21.

Paré, G., Trudel, M.-C., Jaana, M., & Kitsiou, S. (2015). Synthesizing information systems knowledge: A typology of literature reviews. *Information & Management*, *52*(2), 183-199.

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.

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, *63*(10), 1872-1897.

Qureshi, R., Shaughnessy, D., Gill, K. A. R., Robinson, K. A., Li, T., & Agai, E. (2023). Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation? *Systematic Reviews*, *12*(1), 72.

Rahut, S. K., Ullah, M. H., Hafiz, T. B., & Akhter, S. (2022). A recommender engine for scientific paper peer-reviewing system. In M. Virvou, G. A. Tsihrintzis, N. G. Bourbakis, & L. C. Jain (Eds.), *Handbook on artificial intelligence-empowered applied software engineering: Vol. 2: Smart software applications in cyber-physical systems* (pp. 39-53). Springer.

Razack, H. I. A., Mathew, S. T., Saad, F. F. A., & Alqahtani, S. A. (2021). Artificial intelligence-assisted tools for redefining the communication landscape of the scholarly world. *Science Editing*, *8*(2), 134-144.

Resch, K. I., Ernst, E., & Garrow, J. (2000). A randomized controlled study of reviewer bias against an unconventional therapy. *Journal of the Royal Society of Medicine*, *93*(4), 164-167.

Resnik, D. B. (2011). A troubled tradition: It's time to rebuild trust among authors, editors and peer reviewers. *American Scientist*, *99*. https://link.gale.com/apps/doc/A259079742/ITOF?u=monash&sid=bookmark-ITOF&xid=4d72b76f

Resnik, D. B., & Elmore, S. A. (2016). Ensuring the quality, fairness, and integrity of journal peer review: A possible role of editors. *Science and Engineering Ethics*, *22*(1), 169-188.

Resnik, D. B., Gutierrez-Ford, C., & Peddada, S. (2008). Perceptions of ethical problems with scientific journal peer review: An exploratory study. *Science and Engineering Ethics*, *14*(3), 305-310.

Schroter, S., Black, N., Evans, S., Godlee, F., Osorio, L., & Smith, R. (2008). What errors do peer reviewers detect, and does training improve their ability to detect them? *Journal of the Royal Society of Medicine*, *101*(10), 507-514.

Schulz, R., Barnett, A., Bernard, R., Brown, N. J. L., Byrne, J. A., Eckmann, P., Gazda, M. A., Kilicoglu, H., Prager, E. M., Salholz-Hillel, M., ter Riet, G., Vines, T., Vorland, C. J., Zhuang, H., Bandrowski, A., & Weissgerber, T. L. (2022). Is the future of peer review automated? *BMC Research Notes*, *15*(1), Article 203.

Sejnowski, T. J. (2023). Large language models and the reverse Turing test. *Neural Computation*, *35*(3), 309-342.

Shapiro, L., & Spaulding, S. (2021). Embodied cognition. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition ed.). https://plato.stanford.edu/entries/embodied-cognition/

Shneiderman, B. (2020a). Design lessons from AI’s two grand goals: Human emulation and useful applications. *IEEE Transactions on Technology and Society*, *1*(2), 73-82.

Shneiderman, B. (2020b). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human-Computer Interaction*, *36*(6), 495-504.

Sipior, J. C. (2018). Peer reviewer non-performance: The need to replenish the “commons”. *Communications of the Association for Information Systems*, *42*(1), 636-645.

Smith, O. M., Davis, K. L., Pizza, R. B., Waterman, R., Dobson, K. C., Foster, B., Jarvey, J. C., Jones, L. N., Leuenberger, W., Nourn, N., Conway, E. E., Fiser, C. M., Hansen, Z. A., Hristova, A., Mack, C., Saunders, A. N., Utley, O. J., Young, M. L., & Davis, C. L. (2023). Peer review perpetuates barriers for historically excluded groups. *Nature Ecology & Evolution*, *7*(4), 512-523.

Smith, R. (2006). Peer review: A flawed process at the heart of science and journals. *Journal of the Royal Society of Medicine*, *99*(4), 178-182.

Spezi, V., Wakeling, S., Pinfield, S., Fry, J., Creaser, C., & Willett, P. (2018). “Let the community decide”? The vision and reality of soundness-only peer review in open-access mega-journals. *Journal of Documentation*, *74*(1), 137-161.

Spier, R. E. (2002a). The history of the peer-review process. *Trends in Biotechnology*, *20*(8), 357-358.

Spier, R. E. (2002b). Peer review and innovation. *Science and Engineering Ethics*, *8*(1), 99-108.

Spirling, A. (2023). Open generative AI models are a way forward for science. *Nature*, *616*, 413.

Spyns, P., & Vidal, M.-E. (2015). *Scientific peer reviewing: Practical hints and best practices*. Springer.

Stafford, T. (2018). Reviews, reviewers, and reviewing: The “tragedy of the commons” in the scientific publication process. *Communications of the Association for Information Systems*, *42*, 624-629.

Starbuck, W. H. (2005). How much better are the most-prestigious journals? The statistics of academic publication. *Organization Science*, *16*(2), 180-200. https://doi.org/10.1287/orsc.1040.0107

Storey, V. C., Lukyanenko, R., Maass, W., & Parsons, J. (2022). Explainable AI: Opening the black box or Pandora’s Box? *Communications of the ACM*, *65*(4), 27-29.

Straub, D. W. (2008). Type II reviewing errors and the search for exciting papers. *MIS Quarterly*, *32*(2), v-x.

Susarla, A., Gopal, R., Thatcher, J. B., & Sarker, S. (2023). The Janus effect of generative AI: Charting the path for responsible conduct of scholarly activities in Information Systems. *Information Systems Research*, *34*(2), 399-408.

Tang, G. (2023). Letter to editor: Academic journals should clarify the proportion of NLP-generated content in papers. *Accountability in Research*, 1-2. https://doi.org/10.1080/08989621.2023.2180359

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, *185*(4157), 1124-1131.

van Dis, E. A. M., Bollen, J., Zuidema, W., van Rooij, R., & Bockting, C. L. (2023). ChatGPT: Five priorities for research. *Nature*, *614*, 224-226.

Wagner, G., Lukyanenko, R., & Paré, G. (2022). Artificial intelligence and the conduct of literature reviews. *Journal of Information Technology*, *37*(2), 209-226.

Waltman, L., Kaltenbrunner, W., Pinfield, S., & Woods, H. B. (2023). How to improve scientific peer review: Four schools of thought. *Learned Publishing*, *36* (3), 334-347.

Wang, H., Li, J., Wu, H., Hovy, E., & Sun, Y. (2022). Pre-Trained Language Models and Their Applications. *Engineering*. https://doi.org/https://doi.org/10.1016/j.eng.2022.04.024

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). *Emergent abilities of large language models.* ArXiv. https://arxiv.org/abs/2206.07682

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, *9*(4), 625-636.

About the Authors

Ron Weber is an emeritus professor at Monash University and The University of Queensland. His research interests are primarily in the areas of conceptual modeling and the use of philosophical theories to better understand information systems phenomena. He has a BCom. (Hons) degree from The University of Queensland and an MBA and PhD from the University of Minnesota. He is a past dean of the Faculty of Information Technology at Monash University.