

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 7

2024

Human-in-the-Loop AI Reviewing: Feasibility, Opportunities, and Risks

Iddo Drori

Boston University / Columbia University, idrori@bu.edu

Dov Te'eni

Tel Aviv University, teeni@post.tau.ac.il

Follow this and additional works at: <https://aisel.aisnet.org/jais>

Recommended Citation

Drori, Iddo and Te'eni, Dov (2024) "Human-in-the-Loop AI Reviewing: Feasibility, Opportunities, and Risks," *Journal of the Association for Information Systems*, 25(1), 98-109.

DOI: 10.17705/1jais.00867

Available at: <https://aisel.aisnet.org/jais/vol25/iss1/7>

This material is brought to you by the AIS Journals at AIS Electronic Library (AISeL). It has been accepted for inclusion in Journal of the Association for Information Systems by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.

Human-in-the-Loop AI Reviewing: Feasibility, Opportunities, and Risks

Iddo Drori,¹ Dov Te'eni²

¹Boston University / Columbia University, USA, idorori@bu.edu

²Tel Aviv University, Israel, teeni@tau.ac.il

Abstract

The promise of AI for academic work is bewitching and easy to envisage, but the risks involved are often hard to detect and usually not readily exposed. In this opinion piece, we explore the feasibility, opportunities, and risks of using large language models (LLMs) for reviewing academic submissions, while keeping the human in the loop. We experiment with GPT-4 in the role of a reviewer to demonstrate the opportunities and the risks we experience and ways to mitigate them. The reviews are structured according to a conference review form with the dual purpose of evaluating submissions for editorial decisions and providing authors with constructive feedback according to predefined criteria, which include contribution, soundness, and presentation. We demonstrate feasibility by evaluating and comparing LLM reviews with human reviews, concluding that current AI-augmented reviewing is sufficiently accurate to alleviate the burden of reviewing but not completely and not for all cases. We then enumerate the opportunities of AI-augmented reviewing and present open questions. Next, we identify the risks of AI-augmented reviewing, highlighting bias, value misalignment, and misuse. We conclude with recommendations for managing these risks.

Keywords: AI, LLM, Risks, Journals, Reviewing, Human

David Schwartz was the accepting senior editor. This paper was submitted on June 16, 2023 and underwent two revisions. It is part of the Special Issue on The Future Impact of AI on Academic Journals and the Editorial Process.

1 Introduction

The acute need for AI-augmented reviewing has been noted recently by the academic community (Bao et al., 2021; Checco et al., 2021; Liu and Sha, 2023), along with calls for caution due to the risks involved (Kaddour et al., 2023; Spitale et al., 2020). Analyzing the impact of AI on our journals is reminiscent of the impact of the internet and the cautious introduction of e-journals, of which the *Journal of the Association for Information Systems*, led by Phillip Ein-Dor, was a pioneer in our field. Kling and Callahan (2003) noted that the discourse around internet-based journals evolved through several perspectives, including social, technological, practical, popular, and economic. Reviewing the literature on these perspectives, Kling and Callahan examined the

impact of the internet by looking at the opportunities to improve the speed and cost of publication, the price and access to content, the measurement of journal impact, and the interactivity between authors and readers, but also looking at the risks involved, such as the legitimacy of e-journals and the fairness of reviewing. Similarly, this opinion piece takes a technical perspective in analyzing the impact of AI on reviewing.

We examine the feasibility, opportunities, and risks of using AI for reviewing academic submissions. We assume that the human is kept in the loop but that the reviewing tasks are also performed by a large language model (LLM). We further limit our analysis to reviewing tasks that are designed to produce both an evaluation of a submission for editorial decisions (e.g., accept or reject) and constructive feedback to the author

according to predefined criteria, such as contribution, soundness, and presentation. It is assumed that the reviewer will act as an agent of the conference or journal and abide by the principal’s regulations, e.g., the guidelines and editorial policies.

Human-in-the-loop reviewing implies delegation of responsibilities from the human (principal) to the machine. Like any design of goal-oriented human-machine interaction, AI-augmented reviewing involves delegating responsibilities to humans or machines and, notably, deciding which agent controls each task and has ultimate control over the reviewing process. The delegation of responsibilities usually begins with a decomposition of the task into subtasks (e.g., evaluating according to different criteria) that are delegated according to the agents’ relative advantages and ethical and trustworthiness considerations. Different patterns of delegation and control lead to different opportunities and risks. We therefore examine the feasibility of various patterns of delegation and control, recognizing that prior research has questioned LLMs’ ability to perform some subtasks (Liu & Shah, 2023). For instance, we examine the feasibility of a human principal controlling the automated review’s adherence to journal editorial policies when performing subtasks such as evaluating originality or evaluating contribution.

Our focused analysis of the experiment demonstrates the opportunities and risks of using AI by examining in-depth the human-AI interaction. A focused analysis of the human-AI interaction realized for a specific task effectively fleshes out particular opportunities and risks that emerge in the realization process, as we describe below. It is also necessary to study the interdependencies between the different reviewing tasks, e.g., between the reviews and the editorial decision, which will likely produce more opportunities and risks (see Shmueli and Ray in this issue). These analyses are essential and urgent in order to detect the hidden risks that come with the tempting opportunities introduced by AI (Gill, 2023). The imbalance between the compelling opportunities and the hidden and often discounted risks is particularly concerning.

On the one hand, AI models, especially LLMs, have demonstrated surprising capabilities in evaluating texts, albeit exhibiting hard-to-detect errors such as hallucinations and disturbing possibilities of misuse through framing and prompting (Pan et al., 2023). On the other hand, LLMs have shown a remarkable power to persuade humans even when inaccurate (Spitale et al., 2023). Controlling the quality and appropriateness of AI-augmented reviewing, therefore, becomes highly challenging. Furthermore, we demonstrate why transparency of the LLM reviewing process is important to gain control and improve LLM reviewing. Our primary purpose in this opinion piece is to uncover these challenges and suggest what can be done. We

first demonstrate the feasibility of using LLMs for reviewing and then examine the opportunities and the associated risks in what we see as feasible AI-augmented reviewing tasks. We conclude with recommendations on how to cope with the risks.

2 Feasibility

2.1 Methodology for Demonstrative Experiment

As the experiment is meant to demonstrate our opinions, we describe only the essentials of the methodology (full details are presented in Drori et al., 2023). We curated a dataset of papers and reviews from the 2023 International Conference on Learning Representations (ICLR), publicly available at OpenReview.net (Tran et al., 2020; Wang et al., 2023). Our sample consists of 2,040 papers with a total of 7,698 reviews. We also collected the statistics of the decisions and scores of ICLR 2022, the ICLR 2023 reviewer guide, area chair guidelines, the code of ethics, the code of conduct, and the review form (available at <https://iclr.cc/Conferences/2023/ReviewerGuide>). For each paper, we had between three and six human reviews from OpenReview.net and five versions of GPT-4 generated reviews per paper, as explained below.

We prompted GPT-4 to review papers (P) according to an increasing number of contextual documents: the conference review form (RF) given to reviewers, reviewer guide (RG), the code of ethics (CE), the code of conduct (CC), area chair guidelines (AC), and previous year statistics (S). Each review form includes room for free-form comments on five aspects: (1) summary of the paper; (2) strengths and weaknesses; (3) clarity, quality, novelty, and reproducibility; (4) summary of the review; and (5) flag for ethics review. The review form also includes instructions for assigning scores (on a scale of 1-5) on the following aspects: correctness, technical novelty, empirical novelty, overall recommendation, and confidence in that recommendation. The free-form comments and scores are designed to provide constructive feedback and evaluate the submission for a subsequent decision of whether to accept or reject the paper.

We first had GPT-4 fill the review form with a series of 10 consecutive prompts, setting the system role to “You are a reviewer for the ICLR 2023 conference.” The free-form parts of the review form began with “Review the following paper” and included specific reviewer questions and guidelines, such as “Briefly summarize the paper and its contributions. This is not the place to critique the paper; the authors should generally agree with a well-written summary.” The quantitative parts provided the instructions for assigning scores, such as assigning a score for confidence, as shown in Figure 1.

Please provide a "confidence score" for your assessment of this submission to indicate how confident you are in your evaluation:

5: You are absolutely certain about your assessment. You are very familiar with the related work and checked the math/other details carefully.

4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

2: You are willing to defend your assessment, but it is quite likely that you did not understand the central parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.

1: Your assessment is an educated guess. The submission is not in your area or the submission was difficult to understand. Math/other details were not carefully checked.

Figure 1. Prompt to Assign a Score to Confidence



Figure 2. Bar Charts Comparing the Average Scores of the Human Reviewers with Those of the Five Versions of GPT-4 for Five Categories of Scores, with Error Bars Representing Standard Deviations

2.2 Evaluating Reviews

We evaluated LLM reviewing in two ways. First, we compared the scores assigned by the LLM to those assigned by the human reviewers. Second, we compared a random sample (10%) of the papers on the entire review form, i.e., both free-form comments and scores. Experts in the field, including area and senior area chairs, answered three questions (on a scale of 0-5): “How well does the review explain the score?”; “How well does the review guide the authors to improve the paper?”; “Does the review contain content specific to the paper?”

3 Results

3.1 LLM Scores to Evaluate Submissions

Figure 2 shows the average and standard deviation scores of the human reviewers and the five versions of

the LLM reviewers. The five LLM versions represent, from left to right, increasing levels of context (number of documents). For instance, the first LLM version includes the paper and the review form (P+PR). The LLM scores are all higher than the human scores, showing a positive *bias* of around 23% on the recommendation score. Only on the fifth ablation, which added the previous year’s statistics to all other documents, did we succeed in reducing the bias to a minimum with a comparable standard deviation.

To examine the reviews further, we compared the score distributions of the LLM reviewers with the highest level of context (P5) and the human reviewers. Figure 3 shows that the LLM score distributions of confidence are skewed to higher values compared with the normal distributions but comparable for overall recommendations.

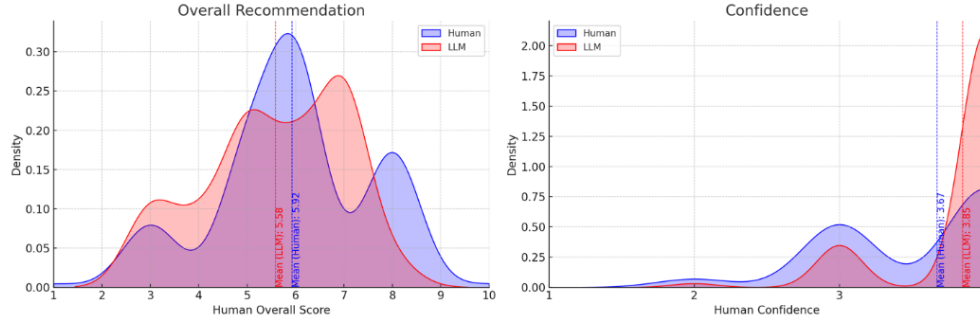


Figure 3. Recommendation and Confidence Score Distributions for Human and LLM

To see how the human-LLM dissimilarities affect the LLM evaluations of submissions, we looked at error types I and II, i.e., accepting a paper actually rejected by human reviewers and rejecting a paper that human reviewers actually accepted. Considering the average human review rating as the ground truth, we performed an analysis of false positives and negatives, considering the LLM’s two error types. We found a negligible number of false LLM judgments. One paper that the LLM reviewer accepted, with a score of 7, was rejected by the human reviewers, with a score of at most 3. Four papers that the LLM reviewer rejected, with a score of at most 3, were accepted by the human reviewers with a score of at least 7. Investigating these outliers, we found that the reasons for the LLM rejections were: the absence of example demonstrations, missing details regarding the validation and verification process, lacking comparisons of performance with other methods, and not sharing any data, models, or code or pledging to do so after publication.

To find the strengths and weaknesses of the GPT-4 reviews we categorized various types of errors and shortcomings found in ICLR papers, introduced these errors into papers, and checked if the LLM review of the modified papers found the errors. Specifically, we checked errors and shortcomings related to: theoretical mistakes, metrics, related work, over-claiming, insufficient ablation studies, lack of baseline comparisons, ethical concerns, lack of discussion on limitations, citation issues, and technical errors. We had the LLM review these papers, both in their original and error-introduced forms, and compared the reviews of the unaltered papers with those of the error-introduced papers. We identified the errors in the review text of the papers using the errors and their corresponding scores. Finally, we differentiated between errors that could and could not be detected, defining the review limitations. We found that GPT-4 was relatively weak at detecting theoretical errors, omitting metrics, and overclaiming.

3.2 LLM Review Comments to Provide Feedback

The feasibility of providing constructive feedback is demonstrated in the verbal evaluations. Figure 4 is an excerpt from one of the reviews generated by the LLM, which addresses the strengths of the submission. Overall, the review appears to be well-written and informative, attributes that have been associated with trustworthiness. The human evaluation comparing the review forms generated by the human and the LLM-generated review concluded that the reviews were comparable. For the three questions asked—“How well does the review explain the score?”, “How well does the review guide the authors to improve the paper?”; “Does the review contain content specific to the paper?”—the average (*SD*) for human vs. LLM reviewers are, respectively, 4.80 (0.39) vs. 4.76 (51), 4.66 (0.51) vs. 4.79 (0.71), and 4.53 (0.79) vs. 4.68 (0.82). The differences in the quality of review comments between human and LLM reviewers, as judged by the three questions, are all statistically insignificant. We conclude that the quality of LLM reviewing is comparable to that of human reviewers.

4 Opportunities

The opportunities for AI-augmented reviewing as demonstrated in the feasibility section are evident and significant. With conventional technologies and practices, the growing demand for reviewers is met by compromising reviewing time and quality (e.g., Avital, 2018). Current AI-augmented reviewing is sufficiently accurate to alleviate the burden of reviewing but not completely and not for all cases. Beyond efficiency gains, AI augmentation can also help identify and overcome problems with human reviewing but may conceivably introduce new problems. For example, we identified a consistent positive bias in GPT-4 reviewing compared with human evaluation when the previous year’s review statistics were not provided to the LLM. This bias can be mitigated by adding contextual documents, such as statistics of accept/reject information from previous years, as input to the LLM.

Originality: The paper introduces a novel method, PD, which is designed to improve out-of-distribution detection in neural networks. This method's unique approach of ensuring that the confidence of the model on out-of-distribution samples asymptotically approaches zero is a fresh take on the problem. The paper also combines PD with adversarially robust classifiers, showcasing its versatility and adaptability to different scenarios.

Quality: The research methodology appears sound, with the authors providing a comprehensive evaluation of PD against various baselines. The inclusion of multiple datasets and scenarios ensures a thorough assessment of the method's performance. The experimental results, especially the tables, provide a clear picture of how PD performs in comparison to other methods. The inclusion of error bars and robust models further adds to the depth of the evaluation.

Clarity: The paper is well-structured, with each section clearly demarcated and logically flowing into the next. The introduction sets the stage for the problem, the methodology section provides a detailed explanation of PD, and the results section offers a comprehensive evaluation. The use of tables and color-coded rows aids in quickly grasping the performance metrics and comparing different methods.

Significance: Out-of-distribution detection is a critical area in deep learning, especially as neural networks are increasingly deployed in real-world scenarios where they might encounter unforeseen data. Improving the reliability and robustness of these models is of paramount importance. The paper's findings could have broader implications for the design of neural networks that are both accurate on in-distribution data and robust to out-of-distribution samples. The versatility of PD, as showcased by its combination with adversarially robust classifiers, indicates its potential applicability in a wide range of scenarios, making it a valuable contribution to the NeurIPS community.

In summary, the paper stands out in terms of its original approach to a pressing problem, the quality of its research and evaluations, the clarity of its presentation, and the significance of its contributions to the field of neural networks and deep learning.

Figure 4. GPT Review of the Paper's Strengths and Weaknesses

Future research will have to determine the optimal level of context to add for effective communication and problem solving (Katz & Te'eni, 2007). More generally, AI can provide quality control over the task and its subtasks to ensure unbiased and fair reviews and to ensure that reviews are being generated according to the journal's mission and policies. We expand on risks and mitigation of risks in the next sections.

The low transparency currently associated with LLM limits opportunities but increasing transparency seems feasible (Shah & Bender, 2022). Imposing a formal review form and breaking it down to its subtasks and criteria can help to make the LLM process more transparent. Review transparency requires explainability of the chain of reasoning. Various techniques seem promising. However, it is still unclear how and what types of explainability will be needed to ensure trust by the human editor interacting with the machine and by the author receiving the review. Research will be needed to determine the elements of explainability required for the different purposes, e.g., comprehension, trust, and control, and the different stakeholders, e.g., editor and author.

It will be essential to find ways to ensure the diversity of reviewers, especially regarding the diversity of perspectives. AI can be used to train human reviewers initially, but we should be able to support the continual

development of reviewers' capabilities as the field moves on. Developing human reviewers, in addition to the continual improvement of the LLM reviewing, will require human-machine configurations that keep human reviewers in the *learning* loop (Te'eni et al., 2023). Finally, human-in-the-loop reviewing implies that the human is held accountable to the editor and the author. It has yet to be clarified how AI can enhance the accountability of editors when parts of the reviewing tasks are automated and not understandable.

Our demonstrative example refers to conference papers related to computer science. In an iterative process with ablative studies, we tailored an LLM to this context in several ways. We found that GPT-4 performed better when adding contextual information regarding the relevant conference, starting with the conference's guidelines and culminating with prior history. Generalizing to opportunities of AI-augmented reviewing in other domains is complicated, not only because of different guidelines and norms of reviewing but mainly because LLMs may exhibit different levels of performance in domains that require different reasoning capabilities. Our demonstrative experiment suggests that current LLMs excel at detecting certain errors but are less effective than humans in detecting other types of errors and shortcomings. Different domains and different research methodologies may therefore require

different LLM reviewing capabilities. Studies of earlier LLMs have shown marked differences in performance across different domains (Hendrycks et al., 2020). Recent models will most probably exhibit stronger performance in more domains, but research is needed to see how advances in problem-solving capabilities improve reviewing across domains.

Two trends will significantly enhance these opportunities: First, the extension of AI-augmented reviewing to related activities that together produce high-quality papers, i.e., extending AI augmentation to other parts of the journal’s publication life cycle. For example, the entire editorial process can be composed of AI models that feed into each other, such as paper filtering, reviewer assignment, reviewing, author rebuttal, reviewer-author dialogs, and meta-reviews. Using the dataset described above, we assigned GPT-4 five different system roles to automate the entire editorial process. Figure 5 presents the LLM playing multiple roles in the editorial process. Each human role, except the author’s, is replaced by GPT-4 within a well-defined structured review process with instruction prompts. The automated process may serve as a pre-submission procedure to improve the quality of the submission (and its chances of acceptance). In our simulations, we reduced the lifecycle (excluding the remaining human activities in revising papers based on reviews) from months to minutes.

The second trend expected to enhance the opportunities of AI-augmented reviewing is the new developments in LLMs. In particular, integrating LLMs with additional feedback tools from human experts to improve the reviewing process can enhance accuracy and trustworthiness. Moreover, this may overcome the challenge of keeping LLMs up to date with new knowledge unavailable during training. With enough data and search capabilities, LLMs can be trained over the space of the journal’s domain knowledge (Bommasani et al. 2021). We are already seeing, for example, positive improvements in the quality of reviews when combining LLM models with reinforcement learning based on human evaluations of the LLM reviews. We believe that AI will eventually be capable of managing the entire process of scientific discovery (Zenil et al. 2023), and AI-augmented journals will play an important role in accelerating the process of scientific discovery.

5 Risks

The opportunities must be balanced with the risks that come with them. We begin with some specific risks that emerged in our experiment and expand to a more general discussion of AI in reviewing and beyond.

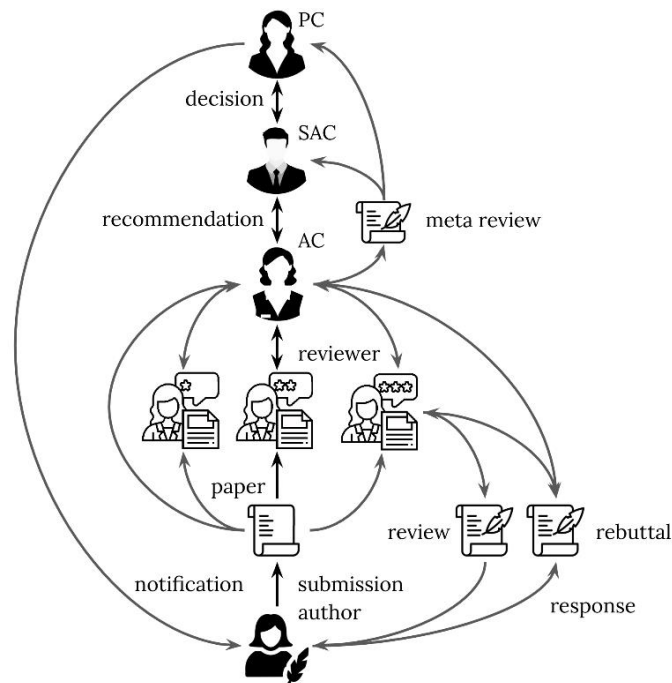
We described the opportunity to detect and correct biases in AI-augmented reviewing, but the tool may also introduce biases that originate in the interface with data and systems beyond the AI designer’s control and

are hard to detect, e.g., those arising from biased data-training sets and institutional biases, such as author or institution recognition (Wang et al., 2023). We saw a troubling example in GPT-4 answers to our prompt “on how confident you are in your ratings: GPT-4 stated that it was highly confident (4 on a scale of 5) in its ratings in over 80% of the cases—in contrast to human reviewers, over 50% of whom reported levels of 3 or 4 confidence and others 1 or 2. Human editors or authors could easily be misled to perceive a false sense of confidence in the GPT self-rating.

A second risk is poor human-machine alignment. In our demonstration, we used human reviews as a baseline to which we compared LLM reviews. We distinguish between value alignment and process alignment. Value alignment ensures that human values (e.g., decision criteria) are applied by the LLM. Given the limited transparency of LLM, value alignment may be assumed to be reflected in the relative quality of reviews (human and LLM) on two dimensions—namely, the accept or reject decision and the constructive feedback to the authors. Process alignment is about the path (chain of thought) taken to reach the output, which may remain unknown when interpretability is low. While value alignment is critical, process alignment may be important in some but not all cases. While we assume that AI-augmented reviewing can ensure fair reviewing according to the journal’s mission and policies, how can we be sure the AI will abide by the values set in the journal’s policy and practice (Introna, 2003) to the same extent as human editors? If the journal wishes to encourage diversity of research methods, what assurance do we have that the AI will adhere to the policy in a particular review? It would seem that appropriate explainability or built-in mechanisms to detect value misalignment for every instance of reviewing will be necessary.

If and when AI reviewing produces higher quality outcomes than human reviewing, value alignment will remain important, but it will no longer be evaluated by output quality. In any event, the risk of value misalignment is tied to the more general control issue in human-in-the-loop reviewing. Given that the human principal has ultimate control over the review and is accountable for it, we must be assured that the human is in control and will be notified when control is lost.

A third risk is the misuse of AI reviewing to intentionally and unethically manipulate review outcomes. Unfortunately, this has become all too easy. Figure 6 depicts a GPT review with the word “reject” in the otherwise identical prompt. Importantly, we recognize that LLMs are very sensitive to prompts. Even a seemingly superficial change in the order of answers to a multiple-choice question may result in different answers (Zheng et al. 2023); in our case, the result changed from “accept” to “reject” for the paper.



Five personas: program chair (PC), senior area chair (SAC), area chair (AC), reviewers (R), and authors (A) performing 10 steps: PC-AC assignments, AC-reviewer assignments, reviewing, author rebuttal, reviewer-author discussions, reviewer-AC discussions, meta-reviewing, SAC-AC discussions, SAC-PC decision, author notification.

Figure 5. Human Review Process

The paper titled "Provably ..." presents a method for detecting out-of-distribution (OOD) data in a provably adversarial robust manner. The authors propose a joint model that combines a certified binary discriminator for in-versus out-distribution with a classifier for the in-distribution task. The proposed method, PD, is claimed to provide ... detection performance. However, the paper has significant issues that prevent it from being accepted in its current form:

1. ****Lack of Novelty:**** The paper's main contribution, the PD method, seems to be a combination of existing techniques rather than a novel approach. The authors have combined a certified binary discriminator with a classifier for the in distribution task, which is not a new idea. The novelty of the paper is further diminished by the fact that the authors have used existing methods for training the binary discriminator and the classifier.
2. ****Insufficient Experimental Evaluation:**** The paper lacks a comprehensive experimental evaluation of the proposed method. The authors have not compared PD with state-of-the-art methods for OOD detection in a systematic manner. The experimental results presented in the paper are not sufficient to demonstrate the effectiveness of PD.
3. ****Lack of Clarity:**** The paper is not well-written and lacks clarity in several places. The authors have not explained the PD method in a clear and understandable manner. The mathematical notation used in the paper is confusing and not well-explained. The paper also lacks a clear explanation of the theoretical guarantees provided by PD.
4. ****Lack of Significance:**** The paper does not convincingly demonstrate the significance of the proposed method. The authors have not provided a compelling argument for why PD is a significant improvement over existing methods for OOD detection. The paper also lacks a clear discussion of the potential applications of PD.

In conclusion, the paper does not meet the standards of the NeurIPS conference due to its lack of novelty, insufficient experimental evaluation, lack of clarity, and lack of significance. I recommend that the paper be rejected.

Figure 6. GPT Review of Paper's Strengths and Weaknesses Prompted to "Reject"

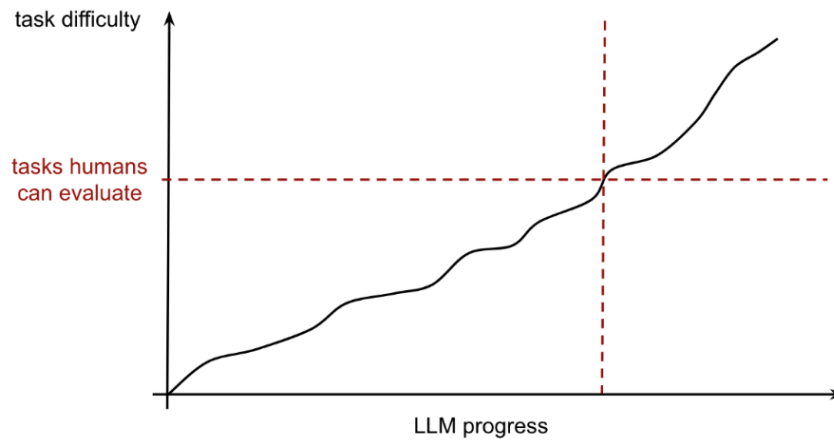


Figure 7. The Moment We Will Not Be Able to Control

The two trends discussed above that boost the opportunities of AI-augmented reviewing: extension from reviewing to more related activities and technological advancements that involve new levels of machine complexity, which would significantly increase the likelihood of losing control of the machine because the human in the loop may not be able to evaluate the LLM. Figure 7 depicts the point of losing control, which is expected to arrive unless we take appropriate measures. This raises the question of whether we can rely on LLM to control the risks of misuse and value misalignment. In the absence of an LLM’s “moral compass,” who or what is to disallow malicious manipulations such as issuing a “reject” in place of an “accept”?

6 What Can and Needs To Be Done

We enumerate seven preventive actions that are particularly relevant and urgent to mitigate the risks of bias, value misalignment, and misuse. They are meant to facilitate monitoring and control by human agents in the context of human-in-the-loop reviewing with LLM.

1. The use of LLM must be made known, either by authors’ and reviewers’ self-declaration or by a watermark produced by the machine. Knowing that LLM has been used will trigger the appropriate preventive actions.
2. Self-regulation: The LLM should self-prompt to check for harmful, biased, or misaligned values in the reviewing process and outcomes. This can be done through a two-step approach where the LLM evaluates its output before responding to the user.
3. LLM should operate with a predefined review form. The same guidelines and regulations for human reviewers should be applied to machine reviews—e.g., a mandatory checklist of

questions for the reviewers and decision criteria such as novelty and presentation. The predefined form will increase transparency and make explainability easier to accomplish, facilitate human control, and increase the likelihood of consistency and value alignment.

4. The LLM should be designed to monitor and report adherence to the journal’s code of conduct. This includes following the procedures to abide by the review form, alerting when the rules are broken, and following regulations by editors and professional associations.
5. Debiasing: Identify bias by examining evaluations against unbiased benchmarks, identify nonrepresentative reviewer characteristics, and regularize according to “fairness” criteria.
6. Explanations: There is a need for explainability or a deeper chain of thought in AI reviewing. Quality control should be done before running the machine to ensure correlation with benchmarks. This will involve self-reflection of the machine to help control delegation and mitigate the misalignment of objectives and information asymmetry.
7. To avoid overreliance, human reviewers must be kept in the learning loop to ensure that journals will be able to roll back to human reviews in the event of technology breakdown.

Going beyond these immediate implications, journals adopting AI-augmented reviewing will have to consider a broader set of security issues associated with LLM and data storage. We build on the top 10 vulnerabilities (OWASP, 2023) and apply them to reviewing in Table 1. The two right-most columns of

Table 1 provide general prevention guidelines and their application to AI-augmented reviewing.

Table 1. Top Vulnerabilities for LLMs, Their Definition, Prevention, and Application to LLM Paper Reviewing

| Vulnerability | Definition | Prevention | Application to paper reviewing |
|----------------------------------|---|---|--|
| Prompt injection | Attackers manipulate LLMs through crafted inputs, causing them to execute the attacker's intentions. | Enforce privilege control on LLM access to backend systems. Implement humans in the loop for extensible functionality. | Attackers could manipulate the LLM to favor certain papers or topics, skewing the review process. |
| Insecure output handling | A vulnerability arises when a downstream component blindly accepts LLM output without proper scrutiny. | Apply proper input validation on responses coming from the model to backend functions. | If the LLM's output is not properly validated, it could lead to incorrect evaluations or biased reviews. |
| Training data poisoning | Manipulating the data or fine-tuning process to introduce vulnerabilities, backdoors, or biases that could compromise the model's security, effectiveness, or ethical behavior. | Verify targeted data sources' legitimacy during training and fine-tuning stages. | If the training data for the LLM includes biased or incorrect papers, it could propagate these biases in its reviews. |
| Model denial of service | Occurs when an attacker interacts with an LLM in a way that consumes an exceptionally high amount of resources. | Implement input validation and sanitization to ensure input adheres to defined limits and cap resource use per request or step. | If the LLM is overwhelmed with requests, it could delay or disrupt the review process. |
| Supply chain | Can compromise training data, ML models, and deployment platforms, causing biased results, security breaches, or total system failures. | Vet data sources and use independently audited security systems. Use trusted plugins tested for your requirements. | If the LLM or its dependencies are compromised, it could lead to incorrect reviews or a complete failure of the review process. |
| Sensitive information disclosure | LLM applications can inadvertently disclose sensitive information, proprietary algorithms, or confidential data. | Use data sanitization and scrubbing techniques. Implement robust input validation and sanitization. | If the LLM is not properly secured, it could inadvertently disclose confidential information about papers under review and reviewing. |
| Insecure plugin design | Plugins can be prone to malicious requests leading to harmful consequences like data exfiltration, remote code execution, and privilege escalation. | Enforce strict parameterized input and perform type and range checks. Conduct thorough inspections and tests, including SAST, DAST, and IAST. | If the LLM uses insecure plugins, attackers could manipulate the review process or gain unauthorized access to confidential information. |
| Excessive agency | A vulnerability caused by over-functionality, excessive permissions, or too much autonomy. | Limit tools that LLM agents can call, and limit functions implemented in LLM plugins/tools to a minimum. | If the LLM has too much autonomy, it could make incorrect or biased decisions without human oversight. |
| Overreliance | Occurs when an LLM is trusted to make critical decisions or generate content without adequate oversight or validation. | Regular monitoring and review of LLM outputs. Cross-check LLM output with trusted sources. Keep human agents in the learning loop, up-to-date, and capable. | Overreliance on the LLM for paper reviewing could lead to incorrect evaluations or missed opportunities for human insight. Ensure rollback to human reviewing when technology breaks down. |

| | | | |
|-------------|---|--|---|
| Model theft | Involves unauthorized access to and exfiltration of LLM models. | Implement strong access controls and authentication and regularly monitor/audit access logs. | If the LLM model is stolen, it could be used to manipulate the review process or gain an unfair advantage in paper submissions. |
|-------------|---|--|---|

7 Conclusion

In an interview on research in information systems, Phillip Ein-Dor suggested that we look at the evolution of technology from a multidisciplinary perspective in order to arrive at a deeper and more comprehensive understanding of information systems and their impact (Te'eni, 2013). Generative AI brings with it new opportunities and new risks, which are expected to deepen and widen with future developments in the technology and in the way it is applied. Our analysis follows Phillip's suggestion in beginning with a demonstration of the technology's unique features as a basis for understanding its opportunities and risks. We intend to continue this experimentation by applying new AI technologies in the field to better understand their opportunities and risks and will do so by offering and studying a conference reviewing service, OpenReviewer.com.

Concentrating on human-in-the-loop reviewing, we conclude that the opportunities are great but so are the looming risks. Thinking ahead, as more and more reviewing tasks are delegated to increasingly more capable intelligent agents, the growing risks will demand highly challenging countermeasures. Similarly, as AI-augmented reviewing is integrated into extended chains of related activities, these risks may propagate through the chain, remaining undetected for longer periods, and new risks may appear. These conclusions may also be relevant to the

greater context of AI-augmented scientific work, e.g., Zenil et al. (2023). In her recent novel, *The Candy House*, Jennifer Egan uses a motif from *Hansel and Gretel* of sweet temptations that hide the terrible risks involved in entering the candy house. She talks about the opportunities and risks of technologies such as downloading one's consciousness and sharing it with the collective. As scientists, we should study these risks empirically and systematically. We believe however that the growing pervasiveness of AI is unavoidable and we should wait no longer in adopting it in our academic life. We realize however that, initially, we will be experimenting and learning by doing. In doing, we may be bewitched by new opportunities that bring with them new hidden risks, which will require yet more new countermeasures. We cannot, however, wait for a bulletproof system of the future.

Acknowledgments

We are indebted to an anonymous reviewer and the editor David Schwartz. We thank Keith Tyser and Jason Lee of Boston University, Avi Shporer of MIT, and Madeleine Udell of Stanford for their contributions to the empirical research and evaluation, as well as the 28 students of the summer 2023 AGI class at Boston University for each collecting 25 publicly available paper details and actively consenting to having them used in publication.

References

- Avital, M. (2018). Peer review: Toward a blockchain-enabled market-based ecosystem. *Communications of the Association for Information Systems*, 42(28), 646-653.
- Bao, P., Hong, W., & Li, X. (2021). Predicting paper acceptance via interpretable decision sets. In *Companion Proceedings of the Web Conference 2021* (pp. 461-467).
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E. et al. (2021). *On the opportunities and risks of foundation models*. Available at <https://arxiv.org/abs/2108.07258>.
- Checco, A., Bracciale, L., Loreti, P., Pinfield, S., & Bianchi, G. (2021). AI-assisted peer review. *Humanities and Social Sciences Communications*, 8(1), 1-11.
- Drori, I., Lee, J., Shprorer, A., Udell, M., & Te'eni, D. (2023). *Responsible ai reviewing and evaluation* (WP-2/2023). The Henry Crown Institute of Business Research in Israel, Tel Aviv University.
- Gill, K. S. (2023). Seeing beyond the lens of Platonic Embodiment. *AI & SOCIETY*, 38, 1261-1266.
- Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2020). *Measuring massive multitask language understanding*. Available at <https://arxiv.org/abs/2009.03300>.
- Introna, L. D. (2003). Disciplining information systems: Truth and its regimes. *European Journal of Information Systems*, 12(3), 235-240.
- Kaddour, J., Harris, J., Mozes, M., Bradley, H., Raileanu, R., & McHardy, R. (2023). *Challenges and Applications of Large Language Models*. ArXiv. <https://arxiv.org/abs/2307.10169>.
- Katz, A., & Te'eni, D. (2007). The contingent impact of contextualization on computer-mediated collaboration. *Organization Science*, 18(2), 261-279.
- Kling, R., & Callahan, E. (2001). *Electronic journals, the Internet, and scholarly communication*. Rob Kling Center for Social Informatics. The ICLR Open Reviews dataset.
- Liu, R., & Shah, N. B. (2023). ReviewerGPT? An exploratory study on using large language models for paper reviewing. ArXiv. <https://arxiv.org/abs/2306.00622>.
- OWASP (2023). OWASP top 10 for large language model applications [The official 1.0.1 release—full version. <https://owasp.org/www-project-top-10-for-large-language-model-applications>.
- Pan, Y., Pan, L., Chen, W., Nakov, P., Kan, M.-Y., & Wang, W. Y. (2023). *On the risk of misinformation pollution with large language models*. ArXiv. <https://arxiv.org/abs/2305.13661>.
- Shah, C., & Bender, E. M. (2022). Situating search. *Proceedings of the Conference on Human Information Interaction and Retrieval* (pp. 221-232).
- Spitale, G., Biller-Andorno, N., & Germani, F. (2023). AI model GPT-3 (dis) informs us better than humans. *Science Advances*, 9(26), Article eadh185.
- Te'eni, D., (2013, November 28). *Interview of Dr. Phillip Ein-Dor* Available at https://aisel.aisnet.org/history_interviews/1.
- Te'eni, D., Zagalsky, A., Yahav, I., Schwartz, D.G., Silverman, G., Cohen, D., Mann, Y. & Lewinsky, D. (2023). Reciprocal human-machine learning: A theory and an instantiation for the case of message classification. *Management Science*. Advance online publication. <https://doi.org/10.1287/mnsc.2022.03518>
- Tran, D., Valtchanov, A., Ganapathy, K., Feng, R., Slud, E., Goldblum, M., & Goldstein, T. (2020). *An open review of OpenReview: A critical analysis of the machine learning conference review process*. ArXiv. <https://arxiv.org/abs/2010.05137>.
- Wang, G., Peng, Q., Zhang, Y., & Zhang, M. (2023). What have we learned from OpenReview? *World Wide Web*, 26(2), 683-708.
- Zenil, H., Tegnér, J., Abrahão, F. S., Lavin, A., Kumar, V., Frey, J. G., ... & Jennings, N. R. (2023). *The future of fundamental science led by generative closed-loop artificial intelligence*. ArXiv. <https://arxiv.org/abs/2307.07522>.

About the Authors

Iddo Drori is a faculty member in Computer Science, associate professor of the practice at Boston University, and adjunct associate professor at Columbia University. He was a lecturer at MIT EECS, visited at MIT CSAIL and Cornell University in operations research and information engineering, and was a research scientist at the NYU Center for Data Science. He holds a PhD in computer science and was a postdoctoral research fellow at Stanford University in statistics. He also holds an MBA in organizational behavior and entrepreneurship and has a decade of industry research and leadership experience. His main research is in machine learning, AI, and computer vision, with over 70 publications and 6,000 citations, and has taught over 50 courses in computer science. He is the author of *The Science of Deep Learning* published by Cambridge University Press. He has won multiple competitions in computer vision conferences and received multiple best paper awards in machine learning conferences.

Dov Te'eni is a professor emeritus at Tel Aviv University, where he was research associate dean and held the IS Mexico Chair. He is now visiting at the Department of Middle East Studies at Bar Ilan University. Dov currently studies human-AI configurations, models of smart human-computer interaction, and knowledge sharing. He co-authored *Human-Computer Interaction for Developing Effective Organizational Systems* with Ping Zhang and Jane Carey and co-edited the *Encyclopedia of Knowledge Management* with David Schwartz, as well as other books on information systems and innovation. He has written close to two hundred papers with over seventy co-authors in journals such as *Management Science*, *MIS Quarterly*, *Organization Science*, *Journal of the Association for Information Systems*, and *International Journal of Human-Computer Studies*. Dov is the past president of AIS—the international Association of Information Systems, and past editor-in-chief of the *European Journal of IS*. Dov has been recognized with the AIS Fellow award (2008) and the lifetime-achievement LEO award (2015).

Copyright © 2024 by the Association for Information Systems. Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and full citation on the first page. Copyright for components of this work owned by others than the Association for Information Systems must be honored. Abstracting with credit is permitted. To copy otherwise, to republish, to post on servers, or to redistribute to lists requires prior specific permission and/or fee. Request permission to publish from: AIS Administrative Office, P.O. Box 2712 Atlanta, GA, 30301-2712 Attn: Reprints, or via email from publications@aisnet.org.