Position: AI Should Verify, Not Judge, Scientific Work

Prabhant Singh [1 2] **Thanh Gia Hieu Khuong*** [3] **Vlasta Sikimić*** [4] **Benedictus Kent Rachmat** [3] **Kola Ayonrinde** [5] **Ihsan Ullah** [6] **Christina Lioma** [7] **Kevin Qinghong Lin** [8] **Luis Oala** [9] **Neil F. Abernethy** [10] **Lele Cao** [11] **Hilde Weerts** [4] **Joaquin Vanschoren** [1]

Abstract

AI reviewers are gaining more attention with the advent of LLMs. With the increasing volume of conference papers, the use of AI-based reviewers has been suggested and implemented to enable faster review cycles, improve review quality, and help the scientific community. **In this position paper, we argue against the replacement of peer review by AI reviewers and advocate the use of AI review tools primarily to verify claims and improve the quality of the scientific work**. We argue that AI review tools should be utilized primarily by authors during manuscript preparation to improve submission quality and streamline downstream evaluation. We ground our argument in the values of peer review and scientific process. Finally, we present research directions for the responsible integration of AI in the peer review process.

1. Introduction

In recent years, scientific venues have faced an unprecedented and exponential increase in submissions (Maslej et al., 2024; 2025). Historically, scientific publishing relied on an implicit equilibrium: the effort required to produce a paper acted as a natural rate-limiting "cost function" that bounded the submission volume (Stephan, 2015). However, the advent of Large Language Models (LLMs) has fundamentally disrupted this equilibrium. As the marginal cost of generating credible scientific claims, methodologies and text approaches zero (Spitzer, 2026), this increase threatens to overwhelm the finite capacity of human peer review (Tran et al., 2020; Liang et al., 2024), in part due to academic pressure to "publish or perish" culture (Beale, 2025).

To accommodate this surge and address the reviewer workload crisis, a number of organizations have turned to AI systems to serve as automated reviewers (Naddaf, 2025), with works advocating for usage of AI to assist in writing reviews (Gruda, 2025). For example, AI reviewers have moved from experimental ideas in research papers to becoming part of the official conference process, with AAAI 2026 (Association for the Advancement of Artificial Intelligence, 2025) having an official AI review, the "La Caixa" Foundation (LCF) implementing the use of AI-based methods for prescreening research proposals, and NeurIPS 2026 piloting an LLM-based reviewer assistant program[1]. In parallel, multiple research works have proposed specialized AI reviewer agents that have been accepted at top conferences (Weng et al., 2025; Zhu et al., 2025b; Zeng et al., 2025; Garg et al., 2025) and commercial tools are being built for review and scientific discovery on top of these works[2]. Research works such as Bauchner and Rivara (2024) advocate that the use of AI as a screening tool prior to external peer review is an inevitable development, while Mann et al. (2025) show that AI reviewer can plausibly enhance error detection and reduce reviewer workload.

In this position paper, **we argue that replacing human peer reviewers with AI reviewers is fundamentally misaligned with the principles and values of the scientific process**. We discuss the importance of peer reviews as a quality filter in scientific venues and the risks of exposing scientific verification to imperfect commercial AI output. We discuss the flaws in current AI-based reviewer evaluations, address the claims around scientific productivity that these tools aim to bridge, and explore why the limitations of AI reviewers are rooted not merely in their current technical capabilities but in how we collectively value and perceive science. For this purpose, it is important to draw a distinction between an AI reviewer and AI review tools. While an *AI reviewer* is defined as a fully automated system that is built to produce a review similar to that of the human, *AI review tools* are

*Equal contribution [1]AMOR/e Lab, Eindhoven University of Technology, Netherlands [2]Microsoft, ISE [3]Université Paris-Saclay, France [4]Eindhoven University of Technology [5]UK AI Security Institute, United Kingdom [6]ChaLearn [7]University of Copenhagen, Denmark [8]University of Oxford, United Kingdom [9]Brickroad Network [10]University of Washington, USA [11]Scholar7, Microsoft Gaming. Correspondence to: Prabhant Singh <p.singh@tue.nl>.

Accepted by ICML 2026 AI for Science Workshop.

[1] https://neurips.cc/Conferences/2026/MainTrackHandbook

[2] https://ai-researcher.net/

systems developed to assess and verify a particular aspect of the scientific publication, such as code, theorems, and grammar. We are critical towards the former but analyze the potential in the latter. Finally, we outline a path forward for **the safe integration of AI-assisted technologies to enhance and broaden peer review**.

2. What are the Goals of Peer Review?

Peer review was first introduced by Henry Oldenburg in 1665 in the *Philosophical Transactions of the Royal Society of London* [3]. The idea of peer review, as the term implies, is for a work to be evaluated by peers within a given scientific field. Peer review acts as a social contract between authors, publishers, and the broader scientific community: an article is published once it is found to be robust and interesting according to the perspective of human experts (peers).

- **Technical Goals**
  - **Quality goal:** Often, the first task in peer review is to determine whether the quality of writing, methods, logic, and internal consistency meet a minimum bar for inclusion in the journal or venue.
  - **Verification goal:** Verification and check of the arguments, methodologies, and data presented in the paper. This ensures that the experiments, methodology supports the main idea of the paper.
  - **Developmental Goal: The "Polishing" of Ideas.** Peer review functions not merely as a gatekeeping mechanism but as a structured, constructive process. Reviewers act as anonymous mentors, providing constructive feedback that pushes authors to clarify their language, strengthen their experiments, and refine their claims.
- **Value driven goals**
  - **Epistemic Value Goal:** Filtering the signal from noise. Deciding if a paper is "interesting" or "significant" requires an understanding of the current zeitgeist of a scientific community.
  - **Ethical Goal:** Peer review often provides science with ethical oversight, especially in studies involving human subjects. For example, the AI community requires an ethical statement in most major conferences, e.g, ICLR: ethical statement; ICML: broader impact statement; NeurIPS: checklist.
  - **Sociocultural Goal:** Establishing scientific truth through collective consensus. Peer review facilitates the negotiation of facts within a community of expert peers, ensuring that knowledge is validated by those within the relevant field.
  - **Alignment Goal:** Peer review also serves as a filter for research that is considered consistent or inconsistent with the accepted scope of the field or publication.
  - **Shared Knowledge Building**: Peer review conducted for specific venues, journals, and conferences serves as a critical mechanism for the collective construction, validation, refinement, and advancement of scientific knowledge.

[3] https://royalsociety.org/journals/publishing-activities/publishing350/history-philosophical-transactions/

Technical Goals (AI-Verifiable)	Value-Driven Goals (Human-Required)
Quality Assurance: Assessing the quality of writing, checking internal consistency, logic, and adherence to minimum methodological standards.	**Epistemic Evaluation:** Deciding a paper's significance and novelty within the current scientific context.
Verification: Auditing mathematical arguments, methods, and data, to ensure that experiments and methods support the paper's claims.	**Alignment & Scope:** Ensuring the research fits the priorities and boundaries of the venue.
Reproducibility: Executing code and validating experimental data pipelines.	**Knowledge Building:** Collaboratively advancing the field's shared understanding.
Scientific Development: Providing constructive feedback to improve clarity, formatting, and presentation.	**Consensus:** Negotiating scientific meaning and validating facts within the expert community.
	Ethical Oversight: Assessing broader societal impacts, dual-use risks, and human subject protections.

Table 1. The demarcation between automated verification and human judgment in peer review.

Peer review serves multiple essential objectives in maintaining the integrity of scientific literature. To better understand how automated systems can assist in this process, we categorize the goals of peer review into two distinct areas: technical goals and value-driven goals, summarized in Table 1.

Technical goals focus on the objective, verifiable components of scientific research. This includes verifying mathematical arguments, ensuring methodological soundness, and confirming the reproducibility of experimental code and data. Moreover, reviewers act as anonymous mentors, providing constructive feedback that pushes authors to clarify their language and refine their claims. Because these tasks rely on formalized rules and internal consistency, they are strong candidates for AI-assisted verification. By automating or semi-automating these checks, the review process

can more efficiently ensure baseline quality and help authors systematically improve and polish their work. For example, an AI system could check if the method section is sufficiently clear and check the validity of statistical claims. It might be better at detecting manipulated data than most human reviewers, and can provide constructive feedback at scale. We elaborate on this in Section 5.

Value-driven goals, in contrast, inherently require human judgment and domain expertise and extend beyond pure verification of the arguments, methodologies, and results to the construction of the scientific facts, ethical integrity, and collective knowledge building. They include determining whether a contribution is scientifically significant or interesting, which requires a deep and practical understanding of the field's current direction. Furthermore, peer review is a collaborative process where experts negotiate scientific consensus, define the boundaries of a venue's scope, and provide critical ethical oversight, particularly regarding broader impacts and human subjects.

We also consider the human editor (i.e., the area or program chair) to be a key figure in the peer-review process who makes the initial and final decisions about a scientific contribution. Both the editor and the reviewers should keep in mind the impact their decisions will have on moral, epistemic, and sociocultural values. When it comes to ethical values, considerations of whether research contributes to the society or if it can be misused.

The sociocultural goal of peer review is to provide a stamp of approval for the research paper and also to facilitate dialog within the community.

The current peer review model operates as a "commons" that relies heavily on scarce human expertise and domain knowledge (Northcraft and Tenbrunsel, 2011). Although an AI reviewer can detect grammatical errors or verify formatting (Korniienko, 2024), it lacks the epistemic grounding to thoroughly negotiate scientific meaning. For example, an AI reviewer would lack access to informal knowledge where researchers collectively interpret emerging directions, reassess what questions matter, and form shared judgments about what counts as timely, surprising, or valuable. If we outsource this validation to an AI reviewer to cope with the volume of submissions, we abandon the social contract of peer review, shifting from a community of practice to a purely transactional mechanism of text generation and automated approval. Thus, in Section 4 we argue that value driven and community goals of peer-review cannot be outsourced even to an otherwise perfect AI reviewer.

3. Do Specialized AI Reviewers Work and Make us Productive?

3.1. Specialized reviewers are developed and evaluated incorrectly

Recently proposed AI review agents such as CycleReviewer (Weng et al., 2025), ReviewRL (Zeng et al., 2025), ScholarPeer (Goyal et al., 2026), DeepReview (Zhu et al., 2025b), and Automated Peer Reviewing (Yu et al., 2024), reveal two methodological shortcomings in their development.

- **Dataset Bias:** The curation of training and evaluation datasets over-indexes heavily on publicly accessible repositories, predominantly OpenReview. This limits the generalizability of these models across diverse scientific disciplines and varied review formats.
- **Misaligned Evaluation Objectives:** Existing AI-based reviewers are predominantly assessed using distance or ranking metrics such as Mean Absolute Error (MAE), mean Squared error (MSE), and Spearman correlation with ground-truth human review scores.

This evaluation creates a critical misalignment: it incentivizes models to mimic human writing styles and surface-level proxies (Checco et al., 2021) rather than satisfying the technical criteria of peer review.

Although achieving high fidelity to human baselines may demonstrate strong language generation capability, matching human scores is an intrinsically flawed objective for scientific evaluation. By optimizing for correlation with human scores and review, they inherit human biases while simultaneously introducing novel machine-specific failure modes. Our claim is further supported by recent studies showing that AI reviewers often overlook counterfactual content (Dycke and Gurevych, 2026) and display a preference for AI-laundered manuscripts compared to human-written ones (Baumann et al., 2026). Furthermore, benchmarking against human scores and review masks a spectrum of systematic vulnerabilities inherent in AI. These include, but are not limited to, systematic bias (Bougie and Watanabe, 2024; Zhu et al., 2025a), adversarial susceptibility (Zhu et al., 2025a; Lin, 2025; Li et al., 2025), novelty blind spots (Akella et al., 2025), and hallucination phenomena (Du et al., 2024).

3.2. AI reviewers do not make reviewing more productive

In this section, we examine common arguments for productivity gains among different stakeholders in the peer review process and provide our critique to those arguments

Reviewer Productivity: A primary motivation for incorporating AI-based reviewers is to reduce the human reviewer

workload.

Critique: With AI-generated reviews, human reviewers are still expected to audit AI output and provide an independent assessment. However, verifying AI-generated claims introduces an additional cognitive load. Furthermore, despite widespread assumptions, no randomized controlled trials (RCTs) currently demonstrate that AI reviewers effectively decrease the total time that human reviewers spend reviewing.

Editor/Program Committee Productivity: Having an additional AI-Review for the editorial team can be helpful when there is a lack of reviewers and provide an additional perspective.

Critique: Having an additional review also involves evaluating that review in the decision making process, which increases effort by the program committee. Having an AI review also opens up the possibility of automation bias within the decision making body.

Author's Productivity: Allowing authors to have a review by an AI-based reviewer before submission can result in better writing, avoiding technical errors, which can save reviewers time and instill confidence in the review process. Providing authors with access to AI-based tools can result in the democratization of resources to improve the writing and development of scientific articles for authors.

Critique: Having an AI reviewer provided to authors can result in papers explicitly crafting their work in an AI-friendly manner. Although not necessarily a bad thing, this could lead to a lack of originality in the authors' writing style and a focus on the areas where AI-based reviewers find it necessary; AI tools have also been shown to expand the impact of scientists, but contracts science's focus (**?**).

Overall, we believe that in the current peer review framework, where stakeholders take full responsibility of their reviews and decisions, the use of AI reviewers is not saving time for reviewers and program committee unless we enable explicit delegation.

3.3. AI reviewers are biased, easily manipulated, and miss critical flaws

Current AI reviewers struggle with foundational evaluation tasks. Although human reviewers are capable of detecting roughly 24-37% of injected manuscript flaws (Shah, 2025), AI reviewers inherently fail to detect faulty reasoning (Dycke and Gurevych, 2025) or recognize the omission of critical information (Fu et al., 2025). Even when explicitly prompted about the presence of flaws, the accuracy of the AI reviewer identification reaches only 39.1% (Xi et al., 2025).

Furthermore, AI reviewers exhibit severe scoring biases. They disproportionately assign highly positive scores (Demetrio et al., 2025; Zhu et al., 2025a) and cannot reliably distinguish between accepted and rejected papers, leaning heavily towards acceptance (Sharma et al., 2026). In contrast, human reviewers show much higher consistency when rejecting papers (Cortes and Lawrence, 2021), despite suffering from overall weak inter-rater agreement ($\kappa = 0.17$) (Bornmann et al., 2010). AI reviewers also demonstrate a distinct impartiality failure, evaluating AI-generated abstracts more favorably than human-written ones (Li et al., 2025; Akpinar et al., 2026) though humans, notably, share this exact preference due to perceived readability (Li et al., 2025; Zhao et al., 2026; Hazra et al., 2026).

Security and reliability present another critical failure point. Unlike human reviewers, AI reviewers can be easily manipulated through adversarial prompt injections hidden in PDFs, extracting favorable scores regardless of scientific merit (Lin, 2025; Zhu et al., 2025a; Baumann et al., 2025). Finally, while authors perceive AI feedback as highly helpful (Liang et al., 2023), this creates an illusion of utility; models fail to mirror critical human observations regarding methodology and context (Suleiman et al., 2024). Automated review tools themselves could be used in a loop with paper-generating LLMs to either train models or select/edit drafts to be scored favorably by the review model. These pieces of evidence indicate that the application of AI reviewers to support peer review remains highly constrained and insufficiently developed.

Critically, AI Reviewers will continue to improve. Several of these cited results are from prior generations of models. The overall performance of several frontier models has continued to advance due to the exploitation of scaling laws, improved inference and grounding, guardrails, and agentic capabilities. We should assume that future AI reviewers will improve and be able to overcome these limitations with better models and evaluation strategies. The AAAI reviewer (Association for the Advancement of Artificial Intelligence, 2025) represents an important first step towards demonstrating that specialized AI reviewers can support some of the technical goals of peer review, although the internals of the system remain private and it has not yet been tested with stress in adversarial settings such as those explored by Dycke and Gurevych (2025), Lin (2025) and Baumann et al. (2025). It is plausible and perhaps inevitable that future AI reviewers will develop robustness against such adversarial attacks and will be able to produce technically sound evaluations.

4. Peer Review with a "Perfect" AI Reviewer

Let us assume the existence of an AI reviewer capable of achieving "perfect" performance defined here as the ability to detect logical flaws, improve prose, and remain robust to

manipulation. How would the peer review process change if a "perfect" AI reviewer existed? More broadly, what would such a system change about how scientific knowledge is evaluated and legitimized? In this thought experiment, the key issue is not only whether AI can judge correctly but whether scientific work should be evaluated by peers or by an automated system.

Fully autonomous scenario

In this scenario, the author's only requirement is to satisfy the AI reviewer. This means that as long as the AI reviewer and the AI meta reviewer give the paper an acceptance, then the paper would be accepted to the venue. We also assume that everything in the paper works and that the paper is well written.

The first risk here is that this scenario makes us too dependent on just the prompting of the AI system. The program chair or organizer of the conference, if any, can simply describe the kind of papers they want, which can tilt the desires of the entire scientific community towards the choices of a few people, restricting the broad spectrum of heterogeneous bottom up ideas that emerge in science. The scenario also questions the meaning of scientific venues; an ArXiv server can simply be connected to this perfect reviewer and output content readers priorities. Ultimately, this scenario undermines the traditional notions of publication, venue, and scholarly validation.

As discussed in Section 2, the peer-review process serves as a "meaning-making" mechanism for a paper. A scientific publication that is evaluated by peers for a venue makes the publication meaningful, and it is the community that makes the venue meaningful. For example, publishing at NeurIPS has value for authors because the scientific community recognizes this venue as a place where especially important contributions are presented. The replacement of the peer reviewers in this scenario risks erosion of the meaning-making aspect and goal of the process.

Another limitation in this scenario is a lack of *Collective knowledge and appreciation*. As discussed in Section 2, AI reviewers lack the framework to incorporate value-driven goals. In peer review, a work is not accepted solely on the basis of technical rigor; there is a sociocultural element to this process that benefits the community. There are situations in which a completely novel application or approach is favored for acceptance by an editor or a group of reviewers, despite other limitations of the paper. These criteria are not possible for an AI reviewer to judge, as currently there is no framework of collective knowledge and community in AI.

5. Where can AI help peer review?

In previous sections, we discussed why an AI reviewer cannot serve as a replacement for traditional peer review. However, we cannot discard the need for AI tools to improve the current peer review process. The question is therefore not whether AI can contribute, but where its contributions are principled, bounded, and verifiable.

This section examines how AI-driven tools deployed as *expert verifiers* rather than reviewer substitutes can support the peer-review process and enhance its overall quality. We organize the discussion around five concrete tasks for which there is growing empirical evidence: (1) reproducibility and results verification, (2) grammar and mechanical checks, (3) mathematical proof verification, (4) citation integrity, and (5) raw experimental data analysis. For each task, we describe the current state of the art, quantify the gap that AI tools can close, and identify the residual limitations that preserve the necessity of human oversight. We have also listed usability of AI in different scenarios in Table 2.

5.1. Reproducibility and Results Verification

Scientific manuscripts may occasionally contain incorrect code, flawed experimental pipelines, or misreported evaluation results. Only about 7% of human reviewers attempt to re-run submitted code due to time and resource constraints (Trisovic et al., 2022). This means that a considerable amount of computational claims in accepted papers are rarely reproduced during review.

Recently, empirical evidence has shown that AI tools can close this reproducibility gap. When prompted to reproduce reported findings from manuscripts alone, LLMs can reproduce approximately 53.2% of results (Dobbins et al., 2025). When deployed, AI agents are able to inspect, modify and execute code in a sandbox environment, AI can reproduce up to 96% of results in certain categories (Shah et al., 2026). Reproducibility frameworks that bundle executable code and data with a manuscript (Meijer et al., 2024) enable AI agents to re-run the full analysis from raw inputs and flag discrepancies for human follow-up. Several benchmarks have also been proposed to assess the capabilities of AI systems for reproducibility of scientific works (Siegel et al., 2024). This suggests that AI tools for reproducibility checks could become an official pre-review step, analogous to automated plagiarism detection.

5.2. Grammar and Mechanical Checks

Grammatical errors, inconsistent notation, and formatting violations are persistent sources of barriers to understanding the scientific content of a paper. AI systems now match human upper bounds of approximately 76% in grammatical error detection (Korniienko, 2024; Bryant and Ng, 2015).

Importantly, this is a domain where the risks of AI involvement are low and the benefits are high.

5.3. Mathematical Proof Verification

In a controlled experiment, peer reviewers often miss algorithmic and mathematical errors, flagging as rarely as 1 in 79 cases in one study (Shah, 2025). This is not a failure of reviewer competence, but of reviewer bandwidth. Checking a multi-page proof in full generality, including all edge cases and boundary conditions, is a task that can take a large proportion of review time while rarely being compensated or recognized. AI combined with formal verification frameworks has demonstrated strong judgment accuracy, exceeding 90% on certain problem classes (Ospanov et al., 2025; Yang et al., 2026). These tools do not yet handle the full breadth of mathematical reasoning encountered in machine learning theory, but they represent a credible path toward formal verification.

5.4. Citation Integrity

Reviewers often lack the ability to verify every reference and, consequently, the integrity of published work is compromised. Mogull (2017) estimate a citation error rate of approximately 14.5% , which means that nearly one in seven cited claims is not actually supported by the cited source. AI systems can now extract, retrieve, and reason over references to achieve up to 97% accuracy in verifying whether a citation genuinely supports the claim it is meant to support (Yuan et al., 2026; noa, 2025).

5.5. Raw Data Analysis

A fifth promising direction is AI-supported analysis of raw experimental data. In current peer review practice, reviewers typically assess processed summaries rather than raw experimental data, operating in an environment of severe information asymmetry. Empirical evidence demonstrates a systemic lack of access to the underlying data and code. For example, in an experiment, among 41 manuscripts in which raw data were explicitly requested, more than 97% failed to provide it, resulting in 21 immediate withdrawals and 19 rejections (Miyakawa, 2020). Similarly, Anderson et al. (2021) evaluated 232 cardiology publications and found that 96.6% lacked access to unmodified datasets, 98.7% omitted step-by-step analysis scripts, and 98.3% withheld complete study protocols. AI tools that allow reviewers to directly interrogate experimental logs and raw datasets rather than rely solely on author-curated tables, figures, or summary statistics can expand the scope of peer review and make overall science more transparent.

Empirical studies show promising evidence that LLMs and multimodal agents are capable of interrogating raw data and verify research integrity at scale. Autonomous AI agents have successfully processed raw proteomics datasets end-to-end to generate expert-validated hypotheses (Ding et al., 2024; Craig and Drăghici, 2024), while machine learning models have achieved an F1 score of 0.92 in detecting contamination artifacts in HPLC traces that are invisible in summary statistics (Gusev et al., 2025). Furthermore, automated pipelines can detect data leakage in benchmarks (Xu et al., 2024) and multimodal frameworks can verify fine-grained alignment between textual claims and visual elements in figures (Shi et al., 2024). By delegating the exhaustive verification of raw data to AI, the peer review process can automatically uncover hidden methodological flaws, ensuring human reviewers can focus their limited bandwidth on the value-driven evaluation of the research.

6. Suggestions and Research Priorities

In this section, we outline suggestions for the safe and fair integration of AI systems into the peer review process, associated risks, and research priorities.

Evaluation with reviewers rather than in isolation: Peer review agents are typically framed as tools to support human reviewers. Although this is a reasonable goal, we advocate that these systems should be assessed in a collaborative setting alongside reviewers. Merely optimizing performance on static peer review benchmark is not a sustainable path to building effective AI reviewers. This also forces us to clarify what we consider a *good review* in a qualitative way. We should conduct RCTs across diverse groups of reviewers to measure the influence of AI reviewers, analogous to Jones (2026), which demonstrates that AI-based reviewer assistants can help make reviews more polite. Related to this suggestions Sikimić (2025) also explores how AI can complement human reviewers in a constructive manner, improving the quality and reliability of the peer review process for grant reviews.

Risk: Human computer/AI interaction research is not easy. There is a natural risk of sampling bias, and defining protocols for these development exercises as well as devising learnings from these studies will be a challenge. These studies are also time consuming, whereas the integration of these systems is already happening.

Allow red teaming: When introducing any kind of AI reviewer in peer review, the AI system should be open for some time to red teaming by the community to find obvious faults. This ensures transparency in the peer review process, improves model quality through public feedback, and demonstrates the limitations of these reviewers in a public manner.

Risk: An AI system may bias topic selection toward what the AI favors or lead to edits that serve the AI rather than human

Task	Human	AI
Grammar / typos	Reliable but slow	Matches human upper bound
Code reproduction	Rarely attempted under time pressure	Reproduces most results with guidance
Proof checking	Often overlooked due to bandwidth	Reliable on bounded problem classes
Citation integrity	Lacks bandwidth to verify every reference	Accurately retrieves and verifies at scale
Raw-data anomaly detection	Limited access; relies on summaries	Can interrogate logs and traces (if available)
Detecting faulty reasoning	Catches a meaningful share	Largely fails, even when prompted
Detecting missing information	Strong on omission	Cannot tell what's not there
Robustness to prompt injection	Resistant	Easily manipulated
Judging significance	Grounded in field zeitgeist	No grounding; mimics human style
Judging novelty	Recognizes timeliness and reframes	Blind spots on genuinely novel work
Ethical assessment	Required by venue norms	Limited to predefined safety guardrails
Community validation	Constitutive of peer review	Cannot participate in social consensus
Final decision	Accountable, contestable	Unaccountable; risks automation bias

Table 2. Comparative strengths and limitations of Peer and AI Review

readers. This risks leading to Goodhart-style[4] overfitting, where the reviewing process measures and rewards not the intended qualitative properties of the paper, but their proxies in the form of topic selection, style of writing or editing.

AI Reviews and feedback should not be provided to the decision making body: AI reviews and any kind of feedback by AI should not be visible to the decision making body, such as the Area-chair or program committee, to avoid automation bias. Multiple studies have shown that even an algorithmic suggestion can exhibit overconfidence in the system and allow for bias from the decision maker, as well as amplify existing biases.

Risk: While this suggestion avoids automation bias, it allows for obvious flaws in the paper detected by AI reviewer to be missed by the program committee (who are also overworked and volunteers and can miss small but very relevant details).

AI as a pre-check mechanism: As pointed out in Section 3.2, current integration of AI tools does not save time, as the reviewer is still responsible for reviewing. Having AI agents for time intensive tasks such as reproducibility check, grammar, and proof checking can serve the review process and improve transparency in science. Having these verified agents as pre-checks can help in reducing errors in proofs, code and allows author to improve the quality of their paper before submitting it to an appropriate venue.

Risk: There are risks regarding the integration of these tools, from financial (who will pay for the tools and model cost) to what kind of proofs can be verified? For example, proof checkers are a great solution for a certain kind of math, like number theory or algebraic geometry, but not for graph theory and category theory. Having precheck requirements poses risks to alienate science which cannot fit in the precheck standards and stick to easy to verify problems. Furthermore, the effectiveness of these tools is hindered by AI's own reproducibility crisis and the worsening transparency in foundation models[5].

[4]In evaluation, Goodhart's Law states that when a metric becomes a target, it ceases to be a reliable measure of success.

7. Alternative Views

In this section, we examine alternative perspectives and counter arguments to our position and offer our critical response to each of them.

We should abolish peer review: Multiple works such as Heesen and Bright (2020) argue for the abolishment of peer review because of its time intensiveness, inefficiency, and bias, and advocate for post publication review instead. Similarly, Mastroianni (2022) argues that peer review does more harm than good because it can give unwarranted credibility to fraudulent papers.

Critique: While this argument raises valid concerns, we should also note that in fields like machine learning and computer science, a large number of papers do have a process similar to post-publication reviews, with papers first being published on ArXiv and then being submitted to a conference or journal. Post-publication review can also create bias in favor of established researchers from top universities vs early career researchers from lesser known institutions. In research areas like medicine, post-publication review can create hype and misinformation around articles with exaggerated claims or misinterpretations of data, which peer review can correct.

AI reviewers will get better and overcome every limitation with new models: AI based reviewers will get better over time and can replace human reviewers once this happens. This claim is valid; we also mention this in our article.

Critique: Critique to this view is addressed in Section 4.

AI review is necessary to scale peer review: Given the

[5]https://crfm.stanford.edu/fmti/December-2025/index.html

exponential growth in submissions, human-only peer review may become fundamentally unsustainable. AI reviewers could act as a necessary scaling mechanism, enabling faster decision cycles and preventing reviewer overload. In this view, even imperfect AI reviewers are preferable to overburdened human reviewers who produce low-quality reviews.

Critique: While scalability is a real concern, replacing human review with AI risks solving a capacity problem by sacrificing epistemic quality and community governance. A more robust solution is to use AI to reduce workload through bounded tasks (e.g., verification, summarization), while preserving human judgment for evaluation and decision-making. Otherwise, we risk creating a high-throughput but low-trust publication system. This view also downplays systematic approaches to deal with scale, such as reciprocal reviewing and a token amount required for excessive number of submissions by authors (IJCAI'25) as well as capping maximum number of submissions per author (KDD).

AI can increase the objectivity of peer review: Human reviewers often disagree about the quality of a submission (Doyle et al., 2015; Fang et al., 2016). This disagreement indicates a lack of objectivity, and a properly designed AI should be able to provide an objective and less biased decision.

Critique: A well-designed and fair AI system could, in principle, produce highly consistent evaluations of scientific submissions. However, this potential advantage introduces two significant concerns. First, because AI systems are easily scalable, a single model could be adopted across many institutions, leading to an algorithmic monoculture. Such homogenization risks standardizing evaluation criteria and publication decisions on a scale, thus reducing intellectual diversity in the scientific literature. Second, even highly capable AI systems remain fallible. If the same model is widely used, important work that it systematically misinterprets or undervalues may struggle to be published at all. For these reasons, consistency alone should not be treated as a virtue, as it may come at the expense of the diversity and plurality that scientific progress depends on.

8. Conclusion

In this position paper, we discuss the capabilities of current AI-based reviewers. **We advocate against using AI review as a replacement for peer review and for using AI tools in peer review for reproducibility, proof verification, citation integrity, and paper improvement**. We grounded our position in the social contract of peer review and advocate that integration of AI review tools should be limited to technical aspects of peer review. We propose guardrails that should be developed during the adoption of these tools. We believe that our work is important and timely in guiding the scientific community in the right direction in developing AI review tools in the future and in integrating AI review tools responsibly.

9. Acknowledgments

We thank Isabelle Guyon for her supervision and contribution for this work.

References

(2025). Developing an AI-Powered Tool for Automatic Citation Validation Using NVIDIA NIM.

Akella, A. P., Siravuri, H. V., and Rohatgi, S. (2025). Prereview to Peer review: Pitfalls of Automating Reviews using Large Language Models. arXiv:2512.22145 [cs].

Akpinar, N.-J., Avula, S., Lee, C. J., Dang, B., Razat, K., and Murdock, V. (2026). LLM or Human? Perceptions of Trust and Information Quality in Research Summaries. arXiv:2601.15556 [cs].

Anderson, J. M., Wright, B., Rauh, S., and Tritz, D. (2021). Evaluation of indicators supporting reproducibility and transparency within cardiology literature. *Heart*, 107(13):1058–1063.

Association for the Advancement of Artificial Intelligence (2025). AAAI launches AI-powered peer-review assessment pilot for AAAI-26. Describes supplemental LLM-generated first-stage reviews and discussion summarization; humans retain decisions.

Bauchner, H. and Rivara, F. P. (2024). Use of artificial intelligence and the future of peer review. *Health Affairs Scholar*, 2(5):qxae058.

Baumann, J., Pei, J., Koyejo, S., and Hovy, D. (2026). Stop automating peer review without rigorous evaluation. In *Post-AGI Science and Society Workshop*.

Baumann, S., Pei, J., Koyejo, O., and Hovy, E. (2025). Stop automating peer review without rigorous evaluation. OpenReview. NeurIPS 2025 position paper on risks of automating peer review.

Beale, R. (2025). In Memorium: The Academic Journal. *Computer*, 58(9):123–126. arXiv:2512.23915 [cs].

Bornmann, L., Mutz, R., and Daniel, H.-D. (2010). A reliability-generalization study of journal peer reviews: A multilevel meta-analysis of inter-rater reliability and its determinants. *PLoS ONE*, 5(12):e14331.

Bougie, N. and Watanabe, N. (2024). Generative Adversarial Reviews: When LLMs Become the Critic. arXiv:2412.10415 [cs].

Bryant, C. and Ng, H. T. (2015). How Far are We from Fully Automatic High Quality Grammatical Error Correction? In Zong, C. and Strube, M., editors, *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 697–707, Beijing, China. Association for Computational Linguistics.

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

Cortes, C. and Lawrence, N. D. (2021). Inconsistency in conference peer review: Revisiting the 2014 NeurIPS experiment.

Craig, D. B. and Drăghici, S. (2024). What's the data say? An LLM-based system for interrogating experimental data. *Bioinformatics*, 40(12). Advance Access 15 November 2024.

Demetrio, L., Apruzzese, G., Grosse, K., Laskov, P., Lupu, E., Rimmer, V., and Widmer, P. (2025). Gen-review: A large-scale dataset of AI-generated (and human-written) peer reviews.

Ding, N., Qu, S., Xie, L., Li, Y., Liu, Z., Zhang, K., Xiong, Y., Zuo, Y., Chen, Z., Hua, E., Lv, X., Sun, Y., Li, Y., Li, D., He, F., and Zhou, B. (2024). Automating exploratory proteomics research via language models. *arXiv preprint arXiv:2411.03743*.

Dobbins, N., Xiong, C., Lan, K., and Yetisgen, M. (2025). Large language model-based agents for automated research reproducibility: An exploratory study in alzheimer's disease.

Doyle, J., Quinn, K., Bodenstein, Y., Wu, C., Danthi, N., and Lauer, M. (2015). Association of percentile ranking with citation impact and productivity in a large cohort of de novo nimh-funded r01 grants. *Molecular psychiatry*, 20(9):1030–1036.

Du, J., Wang, Y., Zhao, W., Deng, Z., Liu, S., Lou, R., Zou, H. P., Venkit, P. N., Zhang, N., Srinath, M., Zhang, H. R., Gupta, V., Li, Y., Li, T., Wang, F., Liu, Q., Liu, T., Gao, P., Xia, C., Xing, C., Cheng, J., Wang, Z., Su, Y., Shah, R. S., Guo, R., Gu, J., Li, H., Wei, K., Wang, Z., Cheng, L., Ranathunga, S., Fang, M., Fu, J., Liu, F., Huang, R., Blanco, E., Cao, Y., Zhang, R., Yu, P. S., and Yin, W. (2024). LLMs Assist NLP Researchers: Critique Paper (Meta-)Reviewing. arXiv:2406.16253 [cs].

Dycke, N. and Gurevych, I. (2025). Automatic reviewers fail to detect faulty reasoning in research papers: A new counterfactual evaluation framework.

Dycke, N. and Gurevych, I. (2026). Automatic reviewers fail to detect faulty reasoning in research papers: A new counterfactual evaluation framework. *Transactions of the Association for Computational Linguistics*.

Fang, F. C., Bowen, A., and Casadevall, A. (2016). Nih peer review percentile scores are poorly predictive of grant productivity. *Elife*, 5:e13323.

Fu, H. Y., Shrivastava, A., Moore, J., West, P., Tan, C., and Holtzman, A. (2025). AbsenceBench: Language models can't tell what's missing.

Garg, M. K., Prasad, T., Singhal, T., Kirtani, C., Mandal, M., and Kumar, D. (2025). ReviewEval: An evaluation framework for AI-generated reviews. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V., editors, *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20542–20564, Suzhou, China. Association for Computational Linguistics.

Goyal, P., Parmar, M., Song, Y., Palangi, H., Pfister, T., and Yoon, J. (2026). Scholarpeer: A context-aware multi-agent framework for automated peer review.

Gruda, D. (2025). Three ai-powered steps to faster, smarter peer review. *Nature*, 10.

Gusev, F., Kline, B. C., Quinn, R., Xu, A., Smith, B., Frezza, B., and Isayev, O. (2025). Machine learning anomaly detection of automated HPLC experiments in the cloud laboratory. *Digital Discovery*, 4:3445–3454.

Hazra, S., Lee, D., Majumder, B. P., and Kumar, S. (2026). Accepted with Minor Revisions: Value of AI-Assisted Scientific Writing. In *Proceedings of the 31st International Conference on Intelligent User Interfaces*, pages 1–22. arXiv:2511.12529 [cs].

Heesen, R. and Bright, L. K. (2020). Is peer review a good idea? *The British Journal for the Philosophy of Science*, page axz029.

Jones, N. (2026). This ai can improve your peer review —and make it more polite. *Nature*, 651(8104):15–16.

Korniienko, O. (2024). Enhancing Grammatical Correctness: The Efficacy of Large Language Models in Error Correction Task.

Li, R., Gu, J.-C., Kung, P.-N., Xia, H., Liu, J., Kong, X., Sui, Z., and Peng, N. (2025). LLM-REVal: Can we trust LLM reviewers yet?

Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D., Yang, X., Vodrahalli, K., He, S., Smith, D., Yin, Y., McFarland, D., and Zou, J. (2023). Can large language models provide useful feedback on research papers? A large-scale empirical analysis. arXiv:2310.01783 [cs].

Liang, W., Zhang, Y., Wu, Z., Lepp, H., Ji, W., Zhao, X., Cao, H., Liu, S., He, S., Huang, Z., Yang, D., Potts, C., Manning, C. D., and Zou, J. Y. (2024). Mapping the Increasing Use of LLMs in Scientific Papers. arXiv:2404.01268 [cs].

Lin, Z. (2025). Hidden prompts in manuscripts exploit AI-assisted peer review.

Mann, S. P., Aboy, M., Seah, J. J., Lin, Z., Luo, X., Rodger, D., Zohny, H., Minssen, T., Savulescu, J., and Earp, B. D. (2025). Ai and the future of academic peer review. *arXiv preprint arXiv:2509.14189*.

Maslej, N., Fattorini, L., Perrault, R., Gil, Y., Parli, V., Kariuki, N., Capstick, E., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., Walsh, T., Hamrah, A., Santarlasci, L., Lotufo, J. B., Rome, A., Shi, A., and Oak, S. (2025). Artificial Intelligence Index Report 2025. arXiv:2504.07139 [cs].

Maslej, N., Fattorini, L., Perrault, R., Parli, V., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., and Clark, J. (2024). Artificial Intelligence Index Report 2024. arXiv:2405.19522 [cs].

Mastroianni, A. (2022). The rise and fall of peer review. *Experimental history*, 14.

Meijer, P., Howard, N., Liang, J., Kelsey, A., and Subramanian, S. (2024). Provide proactive reproducible analysis transparency with every publication. *arXiv preprint arXiv:2408.09103*.

Miyakawa, T. (2020). No raw data, no science: Another possible source of the reproducibility crisis. *Molecular Brain*, 13(1):24.

Mogull, S. A. (2017). Accuracy of cited "facts" in medical research articles: A review of study methodology and recalculation of quotation error rate. *PLoS ONE*, 12(9):e0184727.

Naddaf, M. (2025). More than half of researchers now use ai for peer review — often against guidance. *Nature*. Published online: 15 December 2025.

Northcraft, G. B. and Tenbrunsel, A. E. (2011). Effective Matrices, Decision Frames, and Cooperation in Volunteer Dilemmas: A Theoretical Perspective on Academic Peer Review. *Organization Science*, 22(5):1277–1285.

Ospanov, A., Feng, Z., Sun, J., Bai, H., Shen, X., and Farnia, F. (2025). HERMES: Towards efficient and verifiable mathematical reasoning in LLMs.

Shah, N. B. (2025). Peer review in the era of LLMs: A survey and outlook. Technical report, Carnegie Mellon University. Technical report; periodically updated extended survey.

Shah, S. M. H., Hopfgartner, F., and Bleier, A. (2026). Automating computational reproducibility in social science: Comparing prompt-based and agent-based approaches.

Sharma, V., Joachims, T., and Dean, S. (2026). Do LLMs favor LLMs? quantifying interaction effects in peer review.

Shi, X., Liu, J., Liu, Y., Cheng, Q., and Lu, W. (2024). Every part matters: Integrity verification of scientific figures based on multimodal large language models. *arXiv preprint arXiv:2407.18626*.

Siegel, Z. S., Kapoor, S., Nadgir, N., Stroebl, B., and Narayanan, A. (2024). CORE-bench: Fostering the credibility of published research through a computational reproducibility agent benchmark. *Transactions on Machine Learning Research*.

Sikimić, V. (2025). Fair or flawed? rethinking grant review with generative ai. *Synthese*, 206(6):282.

Spitzer, M. W. H. (2026). The emerging submission crisis in behavioral science. *Trends in Neuroscience and Education*, 42:100276.

Stephan, P. (2015). *How Economics Shapes Science*. Harvard University Press. Google-Books-ID: b5svEAAAQBAJ.

Suleiman, A., von Wedel, D., Munoz-Acuna, R., Redaelli, S., Santarisi, A., Seibold, E.-L., Ratajczak, N., Kato, S., Said, N., Sundar, E., Goodspeed, V., and Schaefer, M. S. (2024). Assessing ChatGPT's ability to emulate human reviewers in scientific research: A descriptive and qualitative approach. *Computer Methods and Programs in Biomedicine*, 254:108313.

Tran, D., Valtchanov, A., Ganapathy, K., Feng, R., Slud, E., Goldblum, M., and Goldstein, T. (2020). An Open Review of OpenReview: A Critical Analysis of the Machine Learning Conference Review Process. arXiv:2010.05137 [cs].

Trisovic, A., Lau, M. K., Pasquier, T., and Crosas, M. (2022). A large-scale study on research code quality and execution. *Scientific Data*, 9(1):60.

Weng, Y., Zhu, M., Bao, G., Zhang, H., Wang, J., Zhang, Y., and Yang, L. (2025). Cycleresearcher: Improving automated research via automated review. In *The Thirteenth International Conference on Learning Representations*.

Xi, S., Rao, V., Payan, J., and Shah, N. B. (2025). FLAWS: A benchmark for error identification and localization in scientific papers.

Xu, R., Wang, Z., Fan, R.-Z., and Liu, P. (2024). Benchmarking benchmark leakage in large language models. *arXiv preprint arXiv:2404.18824*.

Yang, H., Wang, Z., Kang, S., Yang, S., Yu, W., Niu, X., Sun, Y., Hu, Y., Lin, Z., and Zhang, M. (2026). Proof-RM: A scalable and generalizable reward model for math proof.

Yu, J., Ding, Z., Tan, J., Luo, K., Weng, Z., Gong, C., Zeng, L., Cui, R., Han, C., Sun, Q., Wu, Z., Lan, Y., and Li, X. (2024). Automated peer reviewing in paper SEA: Standardization, evaluation, and analysis. In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N., editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10164–10184, Miami, Florida, USA. Association for Computational Linguistics.

Yuan, Z., Shi, K., Zhang, Z., Sun, L., Chawla, N. V., and Ye, Y. (2026). CiteAudit: You cited it, but did you read it? a benchmark for verifying scientific references in the LLM era.

Zeng, S., Tian, K., Zhang, K., Wang, Y., Gao, J., Liu, R., Yang, S., Li, J., Long, X., Ma, J., Qi, B., and Zhou, B. (2025). ReviewRL: Towards automated scientific review with RL. In Christodoulopoulos, C., Chakraborty, T., Rose, C., and Peng, V., editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 16931–16943, Suzhou, China. Association for Computational Linguistics.

Zhao, B., Zhang, J., Whitehouse, C., Jiang, M., Shvartsman, M., Charnalia, A., Magka, D., Shavrina, T., Dunfield, D., Mac Aodha, O., and Bachrach, Y. (2026). APRES: An agentic paper revision and evaluation system.

Zhu, C., Xiong, J., Ma, R., Lu, Z., Liu, Y., and Li, L. (2025a). When your reviewer is an LLM: Biases, divergence, and prompt injection risks in peer review.

Zhu, M., Weng, Y., Yang, L., and Zhang, Y. (2025b). DeepReview: Improving LLM-based paper review with human-like deep thinking process. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T., editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29330–29355, Vienna, Austria. Association for Computational Linguistics.