FEATURES

READY FOR THE ROBOT REVIEWERS?

ILLUSTRATION: STEPHAN SCHMITZ/FOLIOART

As AI's capabilities grow, researchers and publishers are exploring how it can support peer review—and where it still falls short

JEFFREY BRAINARD

You're an editor or a peer reviewer, and unread manuscripts are piling up on your desk. Some of your colleagues are trying out reviewing tools powered by artificial intelligence (AI) that could help relieve the burden. Some tools check for incorrect citations or calculations. Others promise a judgment on whether the manuscript's claims hold up, or even whether they advance knowledge in that field. Would you read those assessments? Trust them? Criticize or reject a paper based on them?

Those questions are looming large for a scientific community that has spent years struggling with an overburdened peer-review system and now faces immense new challenges created by generative AI. The number of scientific publications has doubled in 5 years, to more than 8 million in 2025 (see graphic, p. 967), and some publishers report an even steeper rise in submissions—fueled in part by the wider availability of large language models (LLMs), which can churn out some drafts in minutes. Editors are increasingly struggling to find willing reviewers, contributing to lag times between submission and publication that can drag on for months to more than 1 year. To get just one completed review, editors needed to send an average of 4.5 invitations in 2025—double the number required in 2018, according to a report this year by the Silverchair content-hosting company.

Some researchers and publishers think AI will have to be part of the solution. Automating some of the peer-review process could help speed publication and save researchers' time, advocates say. And, in fact, a growing number of reviewers are employing these tools: In a global survey of 1645 researchers across multiple scientific fields by the publisher Frontiers, just over half of respondents said they used AI for review tasks in 2025, including to help draft comments and in some cases to review methods and statistics. That's up from 29% the previous year—despite many journals prohibiting those uses.

But others are skeptical that machines can give reliable feedback, especially for the most important, core aspects of peer review, including assessing novelty and significance, while avoiding risks such as bias and gaming. "The peer-review crisis is real, and AI looks promising [to help], but irresponsibly rushing towards automating a lot of the judgment might backfire," says Joachim Baumann, a postdoctoral researcher at Stanford University who studies the societal impact of AI. "We need to make sure the tools we deploy are fit for the task."

Research projects and pilots testing AI's capabilities have surged in recent years, aimed at doing just that. "If we are able eventually to get large language models to really do this properly—and increasingly they're obviously getting better—I definitely think they could be a way out of this huge quagmire," says Tom Hope, a researcher at the Allen Institute for AI and the Hebrew University of Jerusalem who is developing AI review tools. As Stephen Turner, a computer scientist at the University of Virginia (UVA), sees it, "The question is not whether AI matches the best human review. It is whether AI-assisted review with clear rubrics outperforms the inconsistent, fatigue-influenced, mood-dependent median review."

For now, most journals are largely barring reviewers from using AI tools until editors determine how they can best support high-quality reviews. Often, however, "People are using these tools in a clandestine, unregulated way," Turner says. "It's kind of the Wild West out there now."

With that in mind, *Science* offers this guide to what researchers are discovering about AI-assisted review and how it might be used productively.

Human reviews have quirks

Evaluations by human reviewers are a vital ingredient in scholarly communication. However, they can be frustratingly subjective and superficial, says Agnieszka Swiatecka-Urban, a physician and researcher in pediatric nephrology at UVA. When humans have flagged deficiencies in her manuscripts, some have written little more than that they "diminish their enthusiasm" for the results. "It's quite irritating because it does not focus on the science," she says. "And there is a temptation—what I see from many reviewers—just to skim through and not be thorough." Human reviewers have also been shown to favor work by prominent researchers and to make harsh, critical comments.

Swiatecka-Urban sees reasons why AI review, if properly designed, can complement or surpass human reviewers. She recently asked AI for suggestions on how to improve several of her manuscripts and a grant application, and found the results useful—for example, in assessing whether the conclusions followed from the results and the same numbers were consistently reported across text, tables, and figures. "What impressed me the most was the objectivity," she says.

Studies bear out that machine-generated reviews may also be more consistent: When both AIs and people were asked to rate conference manuscripts, the machine ratings clustered more tightly than human reviewers', according to a recent study by Baumann and colleagues. Some see this as positive, a correction to inattentive human reviewers. But Baumann warns such clustering could

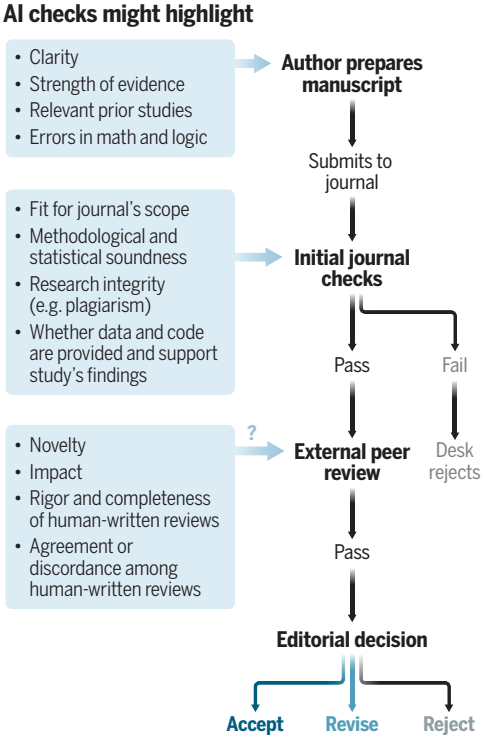
also promote an "intellectual monoculture" that penalizes unorthodox ideas.

Other scholars value the variety of feedback provided by traditional peer review. "What I like about human reviews is that you'll get three different reviews and they're all three completely different—people always pick up on different things," says Ruth Ley, a microbiologist at the Max Planck Institute for Biology. "It just seems sad to lose the heterogeneity of the process."

Reviewers often diverge because of the inherent subjectivity of assessing whether a manuscript's findings are novel—often cited as a key responsibility of peer review, but notoriously thorny to define and pin down. Hope and colleagues are among the teams working to develop AI's capacity to help. Their method uses an LLM to extract key claims about novelty from a manuscript; search online for and summarize related, previously published papers; and evaluate how the new paper departs from that record. In a study of 182 submissions to a recent machine learning conference, human evaluators judged the AI reviews as at least as good at evaluating novelty as human-written ones in most cases. But the approach may miss paradigm-shifting contributions that human experts would recognize and may struggle to keep up with fast-moving fields, Hope acknowledges. Assessing novelty is "a very interesting problem because it's so hard to evaluate," he says.

Where AI could step in

Publishers say artificial intelligence (AI) is not ready to substitute for human assessments of a manuscript's novelty or overall quality, but studies suggest it could inform those judgments.



GRAPHIC: A. MASTIN/*SCIENCE*

What AI peer review does well

One key area in which LLMs may shine is performing routinized, tedious checks for correctness that time-pressed human reviewers often skip. For years, publishers have employed automated checks to make sure manuscripts aren't plagiarized, transparently report methods, and meet other standards. LLMs could bring an expansion of those capabilities, recent research suggests.

In one study, researchers had OpenAI's GPT-5 evaluate 2500 randomly selected papers accepted by three leading machine learning conferences from 2018 to '25. They found the AI was successful at assessing each paper on specific, narrow criteria, such as whether the calculations were correct and the text matched facts in tables, researchers reported in a December 2025 arXiv preprint. The AI found that about one-third of the papers contained at least one error that could affect interpretation, reproducibility, or downstream use—and human evaluators confirmed 83% of those errors, although they found most did not change the paper's main conclusions. The AI was also better at detecting math errors than text mistakes. "Human reviewers may not have the time to go through [a paper] line by line," says study co-author James Zou, an AI researcher at Stanford. "That speaks to a complementarity where AI can help."

Another recent study, reported in May in a preprint on arXiv, undertook a granular, head-to-head comparison of human and AI critiques of the same manuscript. Overall, it found AI reviewers were more diligent than human reviewers at routine but labor-intensive tasks, such as autonomously executing computer code submitted with the manuscript to check whether it supported the paper's claims.

In that study—one of the most extensive of its kind—Ph.D. student Seungone Kim of Carnegie Mellon University and colleagues gathered the preprint versions of 82 papers published in journals in the *Nature* family as well as human-written reviews of those manuscripts, which some of the journals have begun to post with the final published version. They directed three popular LLMs—versions of OpenAI's GPT, Anthropic's Claude, and Google's Gemini—to provide structured reviews of the preprints. The team then asked 45 scientists to score whether each of the 3000 individual critiques in the AI reviews was correct, evidence-based, and significant—meaning it could meaningfully improve the paper—and to identify the best human review for each paper.

Across all papers, GPT scored better than the top-rated human on a composite of the three criteria, and all three LLM models scored better than the lowest rated human. About one-quarter of the AI criticisms were not picked up by any human reviewer and were not trivial, the authors reported. (The AIs also missed the mark in some key areas—see p. 966.)

Some authors are beginning to have access to checkers designed for scientific manuscripts and report finding them useful for improving their drafts prior to submission. This year, three computer science meetings offered authors the opportunity to use Google's Gemini-based Paper Assistant Tool, which inspects methodologies, calculations, and proofs. Among the 1068 authors who used it to prepare for the Conference on Neural Information Processing Systems (NeurIPS), 86% said they found it very or mostly helpful, Google researchers say.

To address concerns that parts of unpublished manuscripts fed through AIs could be used to train an LLM or otherwise become public, the Paper Assistant Tool and other manuscript evaluation services promise to keep the uploaded manuscript in a private repository unconnected to the internet. NeurIPS says it did not show the AI reviews to the conference's referees.

In recent months, the bioRxiv preprint server also began to offer authors the option of sending manuscripts to one of several AI technical-review checkers. The reviews remain confidential to the authors, who can decide whether to revise the manuscript before it appears. Some users report positive experiences, for example with one checker called q.e.d. Science. "I had a more rigorous, attentive, constructive, and useful experience with q.e.d. than with many of my peer reviewers in my 35-plus years submitting papers," says Michael Levin, a developmental biologist and bioengineer at Tufts University who used the service last year.

Where AI reviews fall short

Although Kim's study found AI review has strengths, it also highlighted shortcomings. The LLMs, for example, scored less well overall than human reviewers on the correctness of their assessments. One faulted a paper for not correcting for a known bias in data about air pollution in China—even though the study actually had. Kim's team attributed the mistake to a known limitation of LLMs: Constraints on their working memory may cause them to overlook relevant details, especially when these are spread across multiple sections of a paper and supplementary materials.

The LLMs also displayed a limited grasp of methodological conventions used in certain subdisciplines. For example, a different AI reviewer criticized a particle physics paper produced by a team using the Large Hadron Collider because it did not provide enough detail on the data analysis it used, limiting the study's reproducibility. But such details are routinely housed centrally at the collider's parent lab, CERN, a convention of particle physics the AI did not understand, a human reviewer noted.

In sum, Kim and colleagues write, "AI reviewers could complement, but should not replace, human reviewers. … AI reviewers and human reviewers converge on which parts of a paper warrant review but diverge in how they characterize what they find, meaning that an AI panel is not a drop-in replacement for a human panel."

AI reviews may also make unreasonable requests for additional results. When Ley ran one of her published manuscripts through q.e.d. Science, the AI suggested adding details that would have required time-consuming follow-up studies not necessary to establish the paper's main point, she says. "It brought up things that we know would be the next step, the next paper," she says—but they "would take a ridiculous amount of time. … That's where human judgment would come in."

Some scientists add that AI reviews are verbose and tend to nitpick minor problems while failing to highlight the highest priority issues. Yoshitomo Matsubara, a researcher at the technology company Yahoo, described an AI-written review he saw as part of a pilot for this year's annual meeting of the Association for the Advancement of Artificial Intelligence (AAAI) that tested using AIs to evaluate conference submissions (see p. 967). The review was separated into three sections containing 34 bullet points, totaling 14,000 characters; his own review of the same paper totaled just 4000. "The AI review is too lengthy, and these dots are not well connected," Matsubara says. "It's too much information." And he deemed a summary generated by the AI to be superficial.

The machines can also reflect baked-in biases from the vast body of text on which the LLMs were trained, such as a preference for work by authors from high-prestige institutions. In one experiment, researchers asked an LLM to evaluate multiple versions of papers published in the *Proceedings of the National Academy of Sciences*, identical except for the lead author's name and affiliation, which the team changed. The LLM was more likely to recommend rejection for papers by authors at lower prestige institutions than those at prominent ones, the group led by computational social scientist Anthony Howell of Arizona State University reported in a 2025 arXiv preprint.

Furthermore, AI reviewers struggle to assess results in new fields or narrow subdisciplines in which few prior studies are available to the LLM to inform its judgments. Reflecting AI's tendency toward sycophancy, several studies have found that the machines tend to award scores to manuscripts that are, on average, higher than those from human reviewers. And Baumann's research has shown AI reviewers can be gamed into giving a paper a higher score if authors explicitly direct an LLM to rewrite the paper for that purpose, a trick he calls paper laundering.

At the end of the day, using AI may not even save reviewers time. A research group asked 24 scientists to review two similar published papers selected by the group; the reviewers were directed to use AI assistance for only one of the two reviews. When using AI, the reviewers spent less time on tasks such as checking other relevant literature, but that was offset by the need to check the accuracy of the AI's suggestions, according to a conference presentation last year by a team from University College London (UCL). The net time saved, a few minutes on an hourlong review, was not statistically significant.

Given the limitations of AI, humans need to resist its allure, says Duncan Brumby, a member of that study's research team who studies human-computer interactions. "The more fluent and effortless the [AI] output becomes, the easier it may be to accept something that sounds like a considered review without having done the corresponding intellectual work."

> "We need to make sure the tools we deploy are fit for the task."
>
> **Joachim Baumann**
> Stanford University

ILLUSTRATION: STEPHAN SCHMITZ/FOLIO ART

Conferences and journals are testing the waters

Even amid the uncertainty surrounding AI reviews, and a lack of standards for using them, some journal editors and conferences are dipping their toes into real-world experimentation. When Joydeep Biswas, a researcher in robotics at the University of Texas at Austin, and colleagues were organizing this year's AAAI meeting, they were worried about the rapid growth of submissions and the burden on reviewers. Given the conference's topic, they felt a responsibility and opportunity to study this "socio-technical problem," and whether AI could help ease the load, Biswas says. Still, getting agreement to do the pilot wasn't easy. "I've been on the executive council for a few years and I've never had a discussion which was as contentious and as time-consuming," he says.

They decided their goal would be to test and gather data for how well AI could do the job, while still including ample human input and limiting the extent to which the AI reviews would influence the ultimate decision to accept or reject each submission. They applied a team of LLMs, working together, to help process the 30,000 submissions received, double the number for the previous year's conference. Authors received one AI-written review in addition to at least two human reviews. The human reviews were the only ones that included numerical scores and recommendations about whether the submission should be accepted, and were written without seeing the AI evaluation. Authors then drafted rebuttals to inform final decisions by the conference's human judges.

In a survey of participants, a majority of nearly 6000 authors, reviewers, and judges said they preferred the AI-written reviews over the human ones on six of nine dimensions of quality, such as whether a review was thorough and suggested useful improvements to the research design. Still, authors preferred human reviews over AI in some important categories, including whether the review contained technical errors and overemphasized minor issues. Authors tended to view the AI reviews more favorably than the reviewers and judges. A single AI review "cannot satisfy all three audiences," Biswas says. Authors are eager to improve their paper, whereas judges need a top-line synthesis about whether it makes an original contribution to the field—which often requires intuition only a human can provide. "I came away fully appreciating the importance of human expertise," he says.

"I think there is a way of using humans plus tools so that a human can address a larger number of papers, make better use of their time, and every paper gets some amount of human attention," Biswas adds. He says AAAI plans to repeat the pilot for its 2027 meeting but has no plans to reject papers based solely on AI reviews—a move he expects would draw "stiff opposition" from attendees.

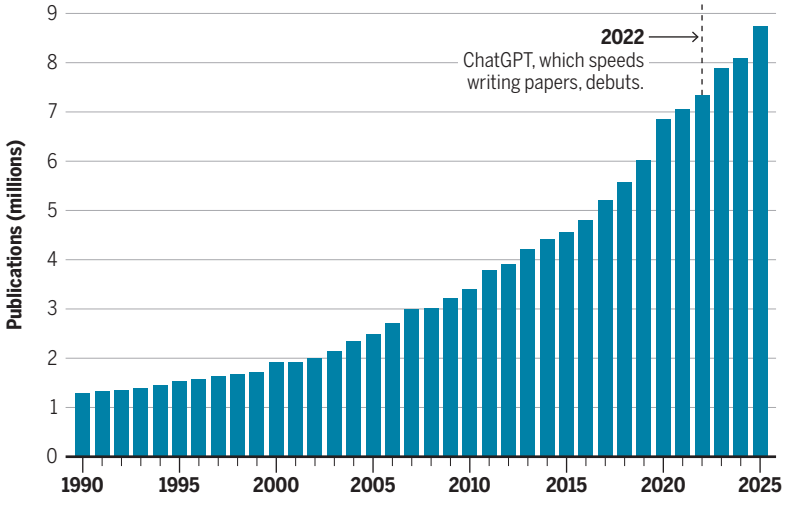
Organizers of another meeting on machine learning, the 2025 International Conference on Learning Representations, ran a trial to see whether LLMs could improve reviews, based on the LLM's analysis of the submitted manuscript and a human-drafted review. After receiving the AI's feedback, 27% of the human reviewers updated their critiques, making them longer on average, and almost all included at least one point from the AI suggestions. A separate panel of human evaluators judged that two-thirds of these revised, AI-informed reviews were better than the original. "We were pleasantly surprised at how many reviewers incorporated the AI feedback, because we know how busy reviewers are," says Zou, senior author of a February paper in *Nature Machine Intelligence* that describes the experiment.

But when the UCL team interviewed reviewers at the same conference, they found mixed views. Some considered the AI critiques to be redundant, or even patronizing because they suggested adding excessive detail, according to the study. "I'm probably confident in my review rather than believing something else [the AI]," one of them told the researchers.

On the journal side, *NEJM AI*, an affiliate of *The New England Journal of Medicine* focused on the use of AI in medicine, began a pilot last year offering a "fast track": an accept-or-reject decision within 7 days based on a combination of AI analysis and review by its human editors. The track is only offered for papers editors think would have a high chance of acceptance through the conventional route. In lieu of sending the paper for outside peer review, an editor performs a full review without AI's help, and other editors debate the human and AI reviews before deciding whether to publish the paper.

Since the journal published the first two papers through this track in November 2025, the quality of the AI analysis has been improving, and so have authors' and editors' comfort levels, says Senior Deputy Editor Arjun Manrai of Harvard Medical School, who studies biomedical informatics. The AI, he says, has been especially helpful in flagging methodological problems and improving papers' clarity. But "we are still verifying the claims they're making and reading them against the paper—trust but verify is still our approach."

"AI will almost certainly have an expanded role [in *NEJM AI*'s reviews] in the future," Manrai says. "In the not too distant past, it would have been quite unimaginable that one of the leading medical AI journals would be using AI in peer review at all. I think the norms are shifting." □



A daunting surge
As generative artificial intelligence accelerates the decades-long increase in scientific publications, editors and reviewers are scrambling to keep up.

Science's AI in Science reporting initiative is supported by Ray Rothrock & family.

CREDITS: (GRAPHIC) V. PENNEY/*SCIENCE*; (DATA) DIMENSIONS

Ready for the robot reviewers?

Jeffrey Brainard

Science **393** (6815), . DOI: 10.1126/science.aem0443