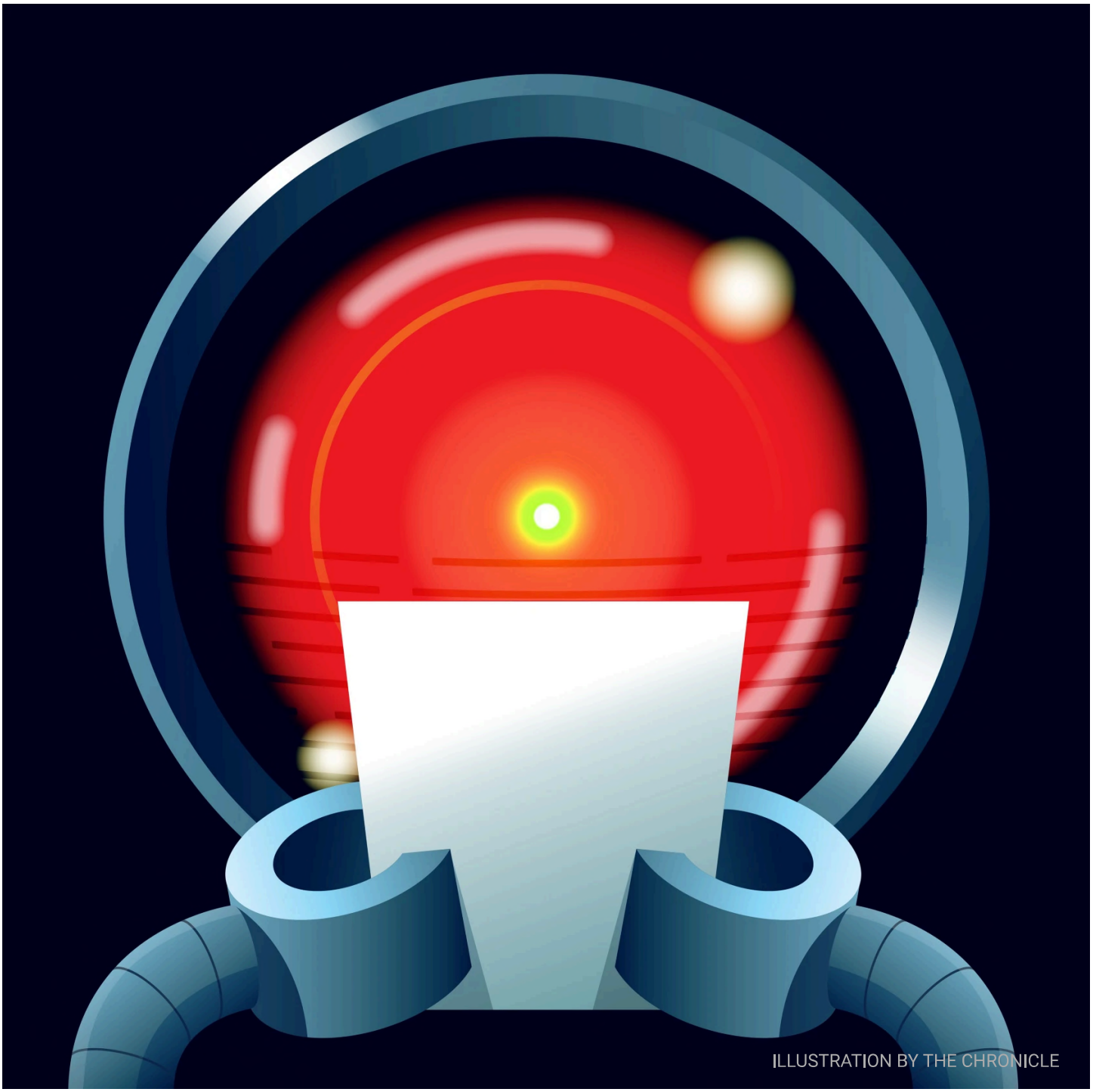
THE CHRONICLE OF HIGHER EDUCATION

AI Scientists Have a Problem: AI Bots Are Reviewing Their Work

ChatGPT is wreaking chaos in the field that birthed it.



HALL OF MIRRORS

By *Stephanie M. Lee*

AUGUST 21, 2024

When Arjun Guha submitted a paper to a conference on artificial intelligence last year, he got feedback that made him roll his eyes. "The document is impeccably articulated," one peer-reviewer wrote, "boasting a lucid narrative complemented by logically sequenced sections and subsections."

Guha, an associate professor of computer science at Northeastern University, knew this "absurd" remark could stem from only one source: an AI chatbot.

"If I wanted to know what ChatGPT thought of our paper," Guha complained on X, "I could have asked myself."

AI is upending peer review, the time-honored tradition in which academics help judge which research should be elevated to publication — and which should go in the reject pile. Under the specter of ChatGPT, no one can be sure anymore that their intellectual labor is being read and judged by humans. Scientists, even those who think generative AI can be a helpful tool, say it's demoralizing to be on the receiving end of an evaluation blatantly outsourced to a robot. And in an ironic twist, this blow to the ego appears to be hitting the AI field most of all: Up to 17 percent of reviews submitted to prestigious AI conferences in the last year were substantially written by large language models (LLMs), a recent study estimated.

Already, there are signs that AI evaluations could be corrupting the integrity of knowledge production. Computer-generated feedback may slightly boost a manuscript's chance of approval, and uploading someone's unpublished data into a chatbot in order to produce a review could amount to a breach of confidentiality policies. These are problems without easy solutions, ones that organizers of computer-science conferences — the main venues for publishing research in that field — are just beginning to acknowledge.

Unfortunately, AI researchers have only themselves to blame.

"That we computer scientists made this thing that is now sort of ruining our review process — yeah, it's quite ironic," Guha said. "It's a bit of a mess."

Even before ChatGPT, peer review was breaking down. Journals and conferences solicit experts to provide anonymous feedback on authors' work, and their opinions are what keep the academic engine churning. Their feedback also helps determine career prospects in a "publish or perish" world. But it's not work most scholars are eager to prioritize, since it's largely unpaid, unrewarded by hiring committees, and time-consuming. Research is also being churned out at record volumes: 2.82 million scientific papers came out in 2022. That's more, one might argue, than humans are actually capable of reading.

Research is being churned out at record volumes — more, one might argue, than humans are actually capable of reading.

Those dynamics are particularly pronounced in the white-hot AI field, which is drawing billions of investment dollars on the premise that it will revolutionize health care, transportation, and every other sector of society. In 2021, 3,014 papers were submitted to the International Conference on Learning Representations (ICLR), a top venue for machine-learning research, and vetted by 4,072 reviewers. This year, the conference reported receiving more than twice as many entries —

7,262 — and accepting about one-third of them. There were also nearly 9,000 reviewers, each of whom evaluated three submissions on average.

"We are absolutely swamped with requests for peer reviews from conferences and journals," said James Zou, an associate professor of biomedical data science at Stanford University. "This is really straining the entire research ecosystem."

Zou said that starting last year, some of his students reported getting peer reviews written in a telltale tone that he described as "more formal" and "a little bit more general." That observation inspired him and a team to analyze reviews submitted to AI conferences before and after the start-up OpenAI released ChatGPT on November 30, 2022. Based on tens of thousands of comments sent to four conferences, they developed an algorithm that can estimate the fraction of substantially AI-modified text with an error rate of less than 2.4 percent.

In a study that has not yet been peer-reviewed or published, they reported that between 6.5 percent and 16.9 percent of the evaluations appeared to contain amounts of AI-generated text that went beyond spell-checking and minor tweaking. In reviews submitted to ICLR, use of adjectives like "commendable," "intricate," and "meticulous" jumped by 9.8-fold, 11.2-fold, and 34.7-fold, respectively, according to the preprint, which was last updated in June.

Bob Carpenter, a senior research scientist at the Flatiron Institute, recently submitted a paper to the Conference on Neural Information Processing Systems, or NeurIPS, a machine-learning conference that last year accepted a quarter of its 13,330 submissions. He and a collaborator disagree on whether one of their evaluations sounded like the handiwork of AI. Carpenter said it was "so vague it's hard to imagine a person writing it" — yet it also contained signs of human error, such as grammar mistakes and logical inconsistencies.

"I think there's a high temptation to use ChatGPT," he said, "just because it's so good."

Its mere existence has injected even more uncertainty into an already-subjective, secretive process, said Mark Dredze, a computer-science professor at the Johns Hopkins University. “If I get a negative review and I suspect it’s written by a language model, and I provide that information to the editor and the paper is still not accepted, was my concern taken seriously? Was it ignored?” he said. “We don’t know.”

AI-generated reviews aren’t exclusive to the AI field. Last year, Andrew D. White, an associate professor of chemical engineering at the University of Rochester, got back a five-sentence review unrelated to the body of a paper he’d submitted to a chemistry journal. He concluded it must have been ChatGPT, which at the time could handle only bits of text. “I think AI tools coming out at the same time that peer review is becoming unsustainable is going to accelerate its demise,” he said.

But compared with other disciplines, AI seems to have a disproportionate number of ChatGPT-flavored reviews. In sharp contrast, Zou’s analysis detected no significant presence of such reviews submitted to *Nature* and its family of journals, which span the natural sciences. It makes sense, Zou said, that the communities developing large language models would be their earliest adopters.

Computer-science conferences also uniquely prioritize peer-reviewing, and speedy reviewing at that. Critiques are openly posted, the process takes place over a few months, and attendees are expected to evaluate work in addition to submitting their own. (Zou’s study found that apparent AI use spiked starting three days before the deadline.) Journals in other fields tend to move more slowly, keep reviews confidential from the public, and be more lenient in letting readers opt out or turn in write-ups late, experts said.

Generative AI tools present other potential downsides for the future of intellectual exchange. Machine-written reviews tend to be more similar to each other than different, Zou and his colleagues found, which could mean that scholars are getting formulaic, less creative responses. With the influx of “noisy” feedback, “a lot of good papers might be rejected from journals or

conferences," Zou said. "Every time a paper is rejected, that also incurs a huge amount of time and cost for the authors."

For this year's ICLR, Guha, the Northeastern computer scientist, turned in a study about how successfully large language models can write code when used by students with little programming experience. Conceptualizing and designing the experiment, running it on dozens of undergraduates across three colleges, and writing up the results took him and his team more than two years.

Last fall, he got back four anonymous reviews, including the one complimenting his "lucid narrative." It declared, too, that "this paper heralds a new dawn for the LLM community" and the analysis was "rendered in an approachable fashion, ensuring it is digestible for a broad readership." (According to the AI-detector GPTZero, there's a 100-percent chance these phrases were pumped out by a large language model.)

"I would put it this way — I *hope* ChatGPT was used," Guha said, "because it would be even more worrying if a person wrote that."

"Peer review is there for a reason. It's a great system; it's super useful. It helps junior researchers a lot to grow, to learn — and if this is not there, I think it's a huge loss."

The numeric scores, however, were less effusive, as were the other reviews. Sensing an imminent rejection, Guha and his colleagues decided to withdraw and submit the paper to another conference, which accepted it. They notified ICLR organizers about the questionable review, but didn't hear back before they pulled out, Guha said. At the time, there was no AI-use policy in place. For its April 2025 gathering in Singapore, ICLR stated for the first time that large language models can be wielded "as a general-purpose assist tool" but are "not eligible for authorship." It also said that "authors and reviewers should understand that they take full responsibility for

the contents written under their name, including content generated by LLMs that could be construed as plagiarism or scientific misconduct."

What troubles Guha is that under the conference's existing ethics code, reviewers are instructed to keep under wraps any confidential information that crosses their desk. ChatGPT trains on user-entered data by default, though some versions of it claim to allow for opting out. "If indeed the author used ChatGPT or something like it, as we suspected, it's quite likely that they violated this policy by uploading our paper, which is supposed to be confidential, to a platform," Guha said.

Programs like ChatGPT could also be making the playing field uneven. According to an analysis of this year's ICLR reviews, a computer-generated score had a 53-percent chance of being higher than a human-given score. (The average difference was about 0.5 points.) And among borderline submissions — meaning their scores didn't make them clear candidates for either acceptance or rejection — those with an AI-generated critique were five percentage points more likely to be accepted than those without. For that analysis, researchers compared the outcomes of pairs that had highly similar topics, reviews, and scores, except that one of the submissions had an AI-written review.

Giuseppe Russo, the preprint's lead author, said that at the beginning of the project, he wasn't necessarily concerned that ChatGPT's footprints were sullying peer review. If AI quickens the articulation of an opinion that ultimately remains the same, "there is no impact on the system," said Russo, a postdoctoral researcher in computer science at the École Polytechnique Fédérale de Lausanne, in Switzerland. But his findings alarmed him. "The fact that it introduces a bias in the system is definitely not positive," he said.

Russo acknowledged that he uses generative AI to help him write reviews — emphasis on help. He said that he always reads the paper and writes his own response, but occasionally asks ChatGPT to analyze it and come up with counterarguments for him to consider incorporating. Similarly, other scientists said

that they value the tool for its ability to distill technical concepts and suggest relevant research to cite.

But Russo also said that it can be deflating to receive a clearly AI-written review, which he believes happened to him once. “If I don’t have someone that actually reads the paper and tells me that I did good work, how can I be sure 100 percent that it’s actually good?” he said. “I think peer review is there for a reason. It’s a great system; it’s super useful. It helps junior researchers a lot to grow, to learn — and if this is not there, I think it’s a huge loss.”

Given how ubiquitous ChatGPT has become in classrooms, tomorrow’s AI scientists may never learn the art of rigorous critique in the first place. Last year, when Jessica Hullman, a professor of computer science at Northwestern University, assigned a class of first-year graduate students to review a paper of their choice, a couple submissions reeked of overly positive language, much to her frustration. “If you’re using the LLM to do the critical thinking for you,” she said by email, “you’ve completely missed the point.”

And of course, scientists are also using ChatGPT to help draft the research they themselves are trying to get published. Russo said that he turns to it mostly to fix the grammar and spelling in his manuscripts, like he does with his peer reviews, but others may be using it for much more. In another recent study by Zou and colleagues, up to 17.5 percent of computer-science preprints released between late 2022 and early 2024 were estimated to be significantly AI-modified. That finding points to an imminent dystopia, one where AI chatbots review AI research produced by AI chatbots.

Last week, an AI start-up in Tokyo announced it had created a program that does exactly that.

In this freewheeling environment, ICLR isn’t the only conference to unveil a policy describing what is and isn’t acceptable. Ahead of its December meeting in Vancouver, Canada, NeurIPS said that authors should disclose

when they use large language models to conduct or analyze their experiments, and that while they can use "any tool they wish for writing the paper, they must ensure that all text is correct and original." It does not have a policy for reviewers, though it has advised organizers to ask "pointed questions to clarify" suspicious-sounding reviews, according to a spokesperson.

Studies have detected AI content in roughly 11 percent to 16 percent of reviews for the most recent ICLR, and 9 percent of those for the most recent NeurIPS. Representatives for the two conferences did not respond to questions about the potential prevalence or influence of AI-written reviews.

For another prominent convening, the Empirical Methods in Natural Language Processing conference, Zou's study estimated that about 17 percent of the reviews submitted last year were heavily AI-written. Thamar Solorio, chair of this year's conference, said that though she had not looked at the paper, this figure seemed "really high." The November meeting, which will be held in Miami, received close to 6,000 submissions, and Solorio said that complaints about potential AI reviews have numbered in the dozens.

The Association for Computational Linguistics, a professional society that sets standards for EMNLP, outlined acceptable uses of generative AI in a policy adopted in June. Reviewers must "read the paper fully and write the content and argument of the review by themselves," can't use a chatbot to write the first draft, and can't upload a manuscript into a "non-privacy preserving" tool, according to the policy. For now, when dubious reviews get flagged, conference organizers have been advised to find replacement reviewers instead of taking action against the reviewers, Solorio said.

The association is now working on creating a committee to investigate the backlog of complaints, according to Solorio, who is a computer-science professor at the Mohamed bin Zayed University of Artificial Intelligence, in the United Arab Emirates, and at the University of Houston. They'll have their work cut out for them, she says, since humans are also imperfect.

In spite of these high-tech headaches, Solorio is an AI optimist. She pointed to ways that natural-language processing could improve peer review: identifying conflicts of interest held by potential reviewers, pointing out mathematical mistakes in manuscripts, detecting plagiarized reviews. And she said that scholars who speak English as a second language, including herself, can benefit from tools that elevate their writing to the fluency level of their native English-speaking colleagues.

"We should set the pace on how to use these tools in peer review in an efficient way," Solorio said of the computer-science community. "We should be doing this, and we should be telling the rest of the scientific world how to use these tools. They're not going to go away."

We welcome your thoughts and questions about this article. Please email the editors or submit a letter for publication.

TECHNOLOGY SCHOLARSHIP & RESEARCH

Stephanie M. Lee

Stephanie M. Lee is a senior writer at *The Chronicle* covering research and society. Follow her on Twitter at @stephaniemlee, or email her at stephanie.lee@chronicle.com.

TOP ARTICLES

SCHOLARSHIP AND RESEARCH
One Scientist Neglected His Grant Reports. Now U.S. Agencies Are Withholding Grants for an Entire University.

RESEARCH INTEGRITY
Stanford Math-Education Expert Has 'Reckless Disregard for

Accuracy,' Complaint Alleges

RESEARCH INTEGRITY

Here's the Unsealed Report Showing How Harvard Concluded That a Dishonesty Expert Committed Misconduct

RESEARCH INTEGRITY

Wanted: Scientific Errors. Cash Reward.