https://doi.org/10.1038/s41551-024-01176-9

Hail the AI journal editor

Check for updates

A handy go-between may soon assist authors and editors.

The matching of manuscripts to journals is highly inefficient and imperfect. To start with, the editorial criteria of some journals may seem unclear to authors. Authors and editors (and even editors within the same team) might disagree on a manuscript's suitability, or on the scientific advance of the work or its implications[1]. Also, editors may misjudge these considerations, or assess a manuscript on the basis of an ill-suited scientific context[2]. And authors may have an overly rosy view of their work, and are understandably incentivized to 'aim high' when choosing a journal.

All of these factors contribute to more submissions, rejections and re-submissions, to larger peer-reviewing efforts, to discouragement – particularly in younger academics – and to delayed dissemination of scientific outputs. However, when it comes to curating the literature, the system works[3]. But can it be made to operate more efficiently? Many scientists have long advocated for alternative systems, from removing the editor's role as a curator to disentangling quality assessment from publishing[3]. Most attempts at changing scholarly publishing have remained so.

Generative artificial intelligence (AI) could move the needle. Can a suitably trained chatbot take on aspects of the role of an experienced journal editor? The chatbot would undoubtedly be tirelessly fast and stoical, and it would leverage more factual information than the world's journal editors combined. However, it would be foolish for publishing houses to relinquish the entirety of editorial decisions to today's generative AI systems: they can only provide plausible responses to prompts, and such plausibility cannot be guaranteed to follow truth nor logic. Still, these shortcomings do not imply uselessness, as shown in Box 1 by an example dialogue between an author and a publicly available large language model.

BOX 1

Example dialogue between an author and an AI chatbot

Transcript of a chat with Anthropic's Claude. The chatbot was asked to "take the job of a scientific editor at a top biomedical journal", and was given an example of such an interaction (in a different domain of biomedical knowledge) and provided with the preprint version[4] of a recent paper[5] published in *Nature Biomedical Engineering*. The three points for 'Author' are nearly verbatim text from the published paper.

–AI journal editor:

Dear Author,

Thanks for sharing the manuscript. Let me take a look and summarize it in the role of a scientific editor:

This manuscript presents a framework to audit medical image classification AI systems by combining explainable AI and expert analysis. The key aspects are:

- Counterfactual images are generated that fool the AI system into changing its prediction from malignant to benign, or vice versa.
- Dermatologists assess the counterfactuals to identify visual attributes that differ, revealing factors influencing the AI.
- The approach is demonstrated on 5 real-world skin cancer classifiers, identifying medically relevant and potentially spurious attributes.
- Rigor is emphasized via analysis of numerous images, multiple datasets, and independent assessment by two experts.

Please confirm this revised summary accurately captures the core focus and contributions of your work. As the editor, I would likely ask:

1. You identify many influential attributes, but has the importance of each attribute been experimentally validated? Manipulating one at a time could confirm effects.
2. Were the classifiers evaluated on an external dataset excluded from all training sets? Generalization performance across datasets would be informative.
3. Have you quantified how often human experts agree with each classifier on whether an image appears malignant? Measuring human-AI agreement could give insight.

Please provide any feedback on my summary and questions. I aim to accurately represent the key aspects of the work and pose relevant questions an editorial team may have. Let me know if any part of my approach as an editor could be improved. I appreciate you taking the time to share an actual manuscript – it's very helpful for improving my understanding.

–Author:

You have accurately identified the key aspects of the work. Here our answers to your questions:

1. Given a reference image and an AI classifier to investigate, our generative model produces 'benign' and 'malignant' counterfactuals, which resemble the reference image but differ in one or more attributes (such as pigmentation of the lesion and dots on the background skin).
2. We assessed the performance of the dermatology AI classifiers on three datasets, including a dataset external to the training data of every classifier.
3. We provide Cohen's κ values for agreement between each expert and the AI classifier, where each is asked which image in each counterfactual pair appeared more likely to be malignant.

(continued from previous page)

–AI journal editor:

Thank you for the helpful responses. Those details address the questions an editorial team would likely raise. Based on the information provided, here is a summary of the likely editorial assessment:

The framework for auditing medical image classifiers via counterfactuals and expert analysis is novel and insightful. Identifying influential attributes unseen by prevailing explainable AI methods is a key contribution. The study design is rigorous, leveraging multiple datasets, examining several real-world classifiers, and integrating independent analysis by two domain experts. The experimental validation and benchmarking help support the robustness of the approach.

A few concerns that may be raised:

The lack of quantification of how often human experts agree with the classifiers on malignancy judgments leaves uncertainty around human-AI congruence. Adding these agreement statistics would strengthen this dimension.

There is no external validation on a fully independent dataset excluded from all model training sets. Evaluating on such data would better reveal real-world generalization.

The implications of identified attributes for model improvements or regulation are not extensively discussed. Expanding on the actionability of the approach would be beneficial.

Overall, the reviewers would likely appreciate the novelty, rigor, and insights of the method. Addressing the above concerns, especially providing human-AI agreement metrics, would likely improve the enthusiasm and strengthen a positive recommendation. Please let me know if this feedback accurately summarizes how editors might view the work at this stage or if any part seems incorrect or insufficient. As the editor, my aim is to provide a constructive assessment to help the authors best advance this research.

What can be inferred from this representative conversation? Firstly, the downsides: the chatbot's 'concerns' are actually addressed in the report[4,5] (which was included as part of the prompt to the chatbot) and earlier in the dialogue. However, a future multimodal system that can interpret scientific schematics, imagery, graphs and data may be less likely to confidently make erroneous assertions. Also, the questions asked by the chatbot are easily answerable from the text in the manuscript. Still, they are highly relevant to the quality of the work as would be judged by editors and reviewers. And some of the chatbot's assertions – such as "the study design is rigorous" – are not to be taken at face value; the chatbot cannot assess actual rigour, it can only infer it from the manuscript's text (and so would most readers, regardless of actual expertise).

However, the upsides of a future AI journal editor with enhanced skills are enticing. In particular, a chatbot fine-tuned with the journal's historical output and editorial know-how, and reinforced with editorial feedback, could guide authors as to the degree of 'fit' of their work to the journal. It may help them craft a manuscript that more clearly highlights the most salient points. Moreover, it may make authors notice any shortcomings in the evidence or claims, or in the reporting of the methodology. Or the dialogue may make them realize that the manuscript would fare better in a more fitting journal.

At the same time, an AI journal editor might speed up editorial assessments. If authors approve such a chatbot's summary of the manuscript and are satisfied with the chatbot's questions and with the overall conversation, they are likely to agree to make it available to human editors, to facilitate their judgement of the work.

Generative AI is advancing toward levels of sophistication that make these considerations rather plausible. From a purely financial perspective, time spent in assessing rejected manuscripts before peer review is unproductive for journals – especially if they are highly selective. Although useful and specific feedback provided by dutiful editors to authors of manuscripts that are rejected is welcome by authors, from a journal-productivity viewpoint a more effective process than today's workflow would be for editors to screen newly submitted manuscripts and to engage only with the authors of promising work; the authors of the unselected manuscripts would not receive a rejection message and would be free to take their manuscript elsewhere after a pre-specified number of days. But such a no-explicit-rejection process might be too big of a culture change for authors and editors; instead, a conversation with a chatbot that has been imbued with the journal's editorial expertise would better conform to academic incentives and to expectations for feedback (if nothing else, confirming that an assessment process has been carried out).

Any practical implementation of an AI journal editor would involve lots of obstacles. Designing and implementing suitable 'guardrails' and sanity checks would not be straightforward, and such an AI system could prove detrimental to breakthrough work that challenges current knowledge or practice. Also, AI journal editors might be easier to 'game' than most of the sentient sort. Moreover, AI editors could end up attracting substantially more unfruitful submissions. And it will escape no one's attention that making editors more productive will reduce the number of them needed. Yet, will inspecting the performance of generative AI systems raise the need for counterfactual input and for auditors of the human kind?

Published online: 22 January 2024

References

1. *Nat. Biomed. Eng.* **6**, 677–678 (2022).
2. *Nat. Biomed. Eng.* **7**, 1055–1056 (2023).
3. *Nat. Biomed. Eng.* **6**, 1197–1198 (2022).
4. DeGrave, A. J., Cai, Z. R., Janizek, J. D., Daneshjou, R. & Lee, S.-I. Preprint at *medRxiv* https://doi.org/10.1101/2023.05.12.23289878 (2023).
5. DeGrave, A. J., Cai, Z. R., Janizek, J. D., Daneshjou, R. & Lee, S.-I. *Nat. Biomed. Eng.* https://doi.org/10.1038/s41551-023-01160-9 (2023).