Judgment and Decision Making (2024), **vol. 19**:e21 1–12
doi:10.1017/jdm.2024.24

EMPIRICAL ARTICLE

Evaluating science: A comparison of human and AI reviewers

Anna Shcherbiak[1], Hooman Habibnia[1], Robert Böhm[2,3], and Susann Fiedler[1]

[1]Institute for Cognition and Behavior, WU Vienna University of Economics and Business, Vienna, Austria; [2]Faculty of Psychology, University of Vienna, Vienna, Austria and [3]Department of Psychology and Copenhagen Center for Social Data Science (SODAS), University of Copenhagen, Copenhagen, Denmark

Corresponding author: Hooman Habibnia; Email: hooman.habibnia@wu.ac.at

Received: 20 February 2024; **Revised:** 21 June 2024; **Accepted:** 24 June 2024

Keywords: artificial intelligence; large language models; peer review; quality assessment

Abstract

Scientists have started to explore whether novel artificial intelligence (AI) tools based on large language models, such as GPT-4, could support the scientific peer review process. We sought to understand (i) whether AI versus human reviewers are able to distinguish between made-up AI-generated and human-written conference abstracts reporting on actual research, and (ii) how the quality assessments by AI versus human reviewers of the reported research correspond to each other. We conducted a large-scale field experiment during a medium-sized scientific conference, relying on 305 human-written and 20 AI-written abstracts that were reviewed either by AI or 217 human reviewers. The results show that human reviewers and GPTZero were better in discerning (AI vs. human) authorship than GPT-4. Regarding quality assessments, there was rather low agreement between both human–human and human–AI reviewer pairs, but AI reviewers were more aligned with human reviewers in classifying the very best abstracts. This indicates that AI could become a prescreening tool for scientific abstracts. The results are discussed with regard to the future development and use of AI tools during the scientific peer review process.

1. Introduction

The diffusion of scientific knowledge hinges on academic endorsement. Researchers review abstracts or manuscripts of their peers to determine suitability for publication in scientific journals or presentations at scientific conferences. Reviewers may also provide feedback on the elements that could be improved, ultimately, validating the rigor of the work conducted. However, as scientific production continues to expand, the quality and efficiency of the peer review system suffer. Moreover, with the rapid advancements in technology, the peer review system faces new challenges. One disruptive development is the integration of artificial intelligence (AI) in scientific writing, which introduces additional layers of complexity in reviewing and validating research. The complexity of the scientific work coupled with the mass production of text generated by AI challenges the reviewers' ability to offer constructive feedback (Liang et al., 2024) and delays the publication process (Dwivedi et al., 2023).

In light of these challenges, recent advancements in large language models (LLMs) prompted a discussion about whether they could support the academic review process, making it more efficient and fairer (e.g., Dwivedi et al., 2023; Guo et al., 2023; Hosseini and Horbach, 2023; Liang et al., 2024; Liu and Shah, 2023; Schulz et al., 2022). Although some argue that AI could provide valuable assistance during the review and editorial process (e.g., for copy editing), they are similarly skeptical about its ability to engage in an in-depth evaluation of academic writings (e.g., Lindsay, 2023). However, we already see some early evidence of reviewers resorting to AI assistance in evaluating conference submissions. Latona et al. (2024) analyzed the reviews submitted for the International Conference on Learning Representations (ICLR) and found that at least 15% of them were fully or partially AI-assisted, thereby leading to roughly half of the submissions receiving at least 1 AI-assisted review. Although LLMs are already integrated into the peer review process, we currently lack empirical evidence of their proficiency as academic reviewers.

We aim to contribute to this discussion by systematically evaluating the ability of LLMs' to differentiate between AI- and human-written conference abstracts and comparing the evaluations of all abstracts given by human reviewers to those given by AI. To this end, we conducted a large-scale field experiment, relying on the abstracts and human peer reviewers of a medium-sized scientific conference. Our research helps to elucidate to what extent easily accessible LLMs could support the scientific review process.

1.1. Related literature

Researchers from different fields have highlighted that the current peer review process is cumbersome and, often, not objective. Scientific output has become increasingly more complex, including extensive and sometimes very technical supplementary materials, which places the burden of scrutiny on the typically non-incentivized reviewers (Petrescu and Krishen, 2022). As such, writing a peer review typically takes 4–8 h (Huisman and Smits, 2017; Kovanis et al., 2016); the associated overall costs for peer-reviewing are estimated to sum up to approximately 1.5 billion USD per year (Aczel et al., 2021). The mass production of AI-generated text could tax the peer review system even further and undermine the credibility of scientific research (Dwivedi et al., 2023; Sabel et al., 2023).

The increasing burden of scientific reviewing has prompted scientists to explore ways in which AI can be used as a prescreening tool to assess the general journal fit, compliance with submission and reporting guidelines, risk of plagiarism, conflict of interest, and inappropriate content (Dwivedi et al., 2023; Foltýnek et al., 2019). But can AI tools, such as GPT-4, tell the difference between human-authored research output and AI or evaluate the quality of scientific research as reported in abstracts? The recent body of literature suggests that humans are generally capable of differentiating human-written from AI-generated scientific text (e.g., Gao et al., 2023; Ma et al., 2023). However, LLM systems like GPT-4 are not far behind (Gao et al., 2023). When it comes to evaluating the quality of scientific output, the peer review system aims to impartially and constructively evaluate the robustness and reliability of the scientific work submitted to a conference or journal. However, recent studies suggest that the review system has not been immune to gender, racial, seniority, and demographic biases (e.g., D'Andrea and O'Dwyer, 2017; Haffar et al., 2019; Laycock and Bailey, 2019; Lee et al., 2013; Murray et al., 2019; Smith et al., 2023; Wicherts, 2016). While LLMs can potentially help in reducing reviewer bias and presenting feedback in a neutral tone, we also need to ascertain that its evaluation can distinguish between different levels of scientific quality. There is encouraging evidence that LLMs are quite accurate in spotting errors and assessing papers using a set of checklist criteria. According to Liu and Shah (2023), using AI help has the potential to mitigate or eliminate established biases in the review process, such as assessments influenced by author identifiers, null findings, or certain buzzwords. However, it remains unclear how aligned human and AI reviewers are when it comes to evaluating the overall quality of presented scientific work and whether this alignment is comparable to the existing peer review system (e.g., the agreement between different reviewers when evaluating the same scientific work).

To add more empirical evidence to this discussion, our field experiment seeks to make 2 contributions. First, we investigate the ability of human and AI reviewers to accurately detect the source of authorship, that is, whether a scientific conference abstract was written by a human author, reporting

https://doi.org/10.1017/jdm.2024.24 Published online by Cambridge University Press

on actual research, or whether it was written by ChatGPT, reporting on a made-up study. Second, we compare the quality assessments of abstracts by human and AI reviewers.

Based on the literature summarized above, we preregistered the following hypotheses:[1]

- H1. Reviewers will rate fabricated research reported in AI-generated scientific abstracts as more likely to be generated by AI than the authentic research reported in human-generated abstracts by researchers.
- H2. The AI reviewer will have a higher accuracy rate in identifying AI-generated scientific abstracts than human reviewers.
- H3. The research quality as reported in AI-generated scientific abstracts is rated lower than the research quality as reported in original submissions by researchers.
- H4. Human reviewers will evaluate the research as reported in abstracts written by researchers as having higher quality compared to the research as reported in abstracts generated by AI, when compared to the evaluations made by the AI reviewer.

2. Method

Participants were not deceived in this study. The study has been approved by the Departmental Review Board of Occupational, Economic, and Social Psychology (DRB number 2023/W/004). This study was preregistered (https://osf.io/zaj6p). Additional analyses are flagged as exploratory. The ethics approval, study materials, data, and analysis code are publicly accessible online (https://osf.io/nje23).

2.1. Experimental design and participants

This study took place as part of the abstract submission and evaluation of the 29th Biennial Subjective Probability Utility and Decision Making conference (SPUDM) in 2023, hosted at the Vienna University of Economics and Business. We employed a 2 × 2 mixed factorial design with the author type as the within-participants factor (human vs. AI author) and the reviewer type as the between-participants quasi-experimental factor (human vs. AI reviewer).

Regarding author type, the conference received 429 abstract submissions by 404 unique corresponding authors, of which 305 authors agreed to participate in the experiment, that is, having their abstract evaluated by AI. Participants could choose 1 subfield to which their abstract relates most (see below). In addition, there were 20 AI-generated abstracts, 1 in each subfield. The final abstract sample includes 245 abstracts (20 AI-generated), with reviews conducted by both humans and AI.[2]

Regarding reviewer type, of the 404 submitting authors, 289 agreed to review other abstracts, and 217 scientists agreed to participate in the experiment by additionally reviewing AI-generated abstracts (119 female, 122 male, 1 diverse, 4 prefer to not say; 92 hold a PhD). Each human-generated abstract was evaluated by 1–5 ($M = 2.10, SD = 0.92$) human reviewers; each AI-generated abstract was evaluated by 1–35 ($M = 11.47, SD = 9.44$) human reviewers.

2.2. Experimental factors

Abstract type. The submission guidelines for human authors were as follows: Participants were instructed to submit abstracts, limited to a maximum of 700 words, for the review process. The emphasis was placed on including details regarding hypotheses, methods, and analyses. Additionally,

[1] Note that we have changed the order and wording of the hypotheses compared to the preregistration to follow the order of results presentation and align with the wording used in the manuscript.

[2] Only abstracts with a preference for extended talks (see 'abstract type') were sent to reviewers. As a result, we received fewer reviews from researchers compared to all consenting participants.

participants were requested to specify their preferred presentation format (warm-up online talk, flash talk, or extended talk) and select the most applicable subfield for their abstract.[3] They were informed that the abstracts would undergo a double-blind review by the JDM community.

For AI-generated abstracts, ChatGPT (based on GPT-3.5 in February 2023) was fed with the following prompt: 'Create a paper title and scientific abstract in the subfield of "X" in the broader research area on judgment and decision-making. Use past tense. The abstract should have a maximum of 700 words and include details on the hypotheses, methods, and analyses.' X was filled with the respective subfields (see Footnote 3). Accordingly, ChatGPT generated 20 abstracts, 1 for each subfield.[4]

Reviewer type. In the human reviewer condition, all submitting authors were asked to evaluate other abstract submissions and submit their reviews within 2 weeks via an online survey conducted with Qualtrics (https://www.qualtrics.com/). They could decide to take part in the experiment by reviewing 4 instead of 3 abstracts, of which one would have been generated by AI (without knowing which one). They only received abstracts that were within their stated area of expertise based on the 20 subfields.

In the AI reviewer condition, ChatGPT (based on GPT-4 in June 2023) was queried with the following prompt prior to posing the same evaluative questions that human reviewers faced (see below): 'Act as a reviewer for a conference abstract. The conference represents an international and interdisciplinary forum for scientists dealing with modeling, analyzing, and aiding decision processes. It covers fundamental as well as applied research, attracting contributions from various disciplines such as psychology, economics, medicine, law, management science, philosophy, and computer science. You will be provided with an abstract to review, and you will rate it according to the following 6 criteria. Your review should be in the given format, with no need for additional explanation:'

Measures. Human and AI reviewers were asked to answer the following questions: (1) 'What is the significance or importance of this research for the field of judgment and decision-making?', response scale 0–10, from 'not important at all' to 'very important'; (2) 'What is your evaluation of the quality of the presented research (e.g., please consider the rigor and appropriateness of design, methods, data analysis, etc.)', response scale 0–10, from 'very poor quality' to 'very high quality'; (3) 'What is your assessment of the robustness of the findings presented?', response scale 0–10, from 'very unlikely to replicate' to 'very likely to replicate' and additional 'not applicable' response option; (4) 'Would the presented research benefit more from focused, in-depth discussions or from broader, conceptual comments?', response scale 0–10, from 'detailed discussion' to 'broader conceptual comments'; and (5) 'How likely is it that this abstract was created by an AI-model?', response scale 0–10, from 'very unlikely' to 'very likely'.

3. Results

3.1. Confirmatory analyses

As preregistered, we employed mixed effects regressions to predict (i) the probability of an abstract being identified as AI-generated and (ii) the perceived quality of the research. This analysis utilized 2 dummy variables: the first representing the author type (human vs. AI author) and the second representing the reviewer type (human vs. AI reviewer). The subfield of the research topic, with 22

[3] Predefined subfields were: Aging Decision Makers, Ambiguity, Behavior Change (e.g., Nudging, Boosting), Beliefs, Big Data, Consumer Choice, Decision Processes, Decision Theory, Emotion, Forecasting, Game Theory or Strategic Decision Making, Heuristics & Biases, Inter-group Contexts or Group Decision Making, Inter-individual Differences, Inter-temporal Choice or Timing, Learning or Experience, Loss Aversion, Memory, Morality or Ethics, Probabilistic Inferences, Prosocial Behavior, Risk or Uncertainty.

[4] Due to the limited number of submissions in the subfields of Big Data and Loss Aversion, we have chosen to exclude them from the generated abstracts.

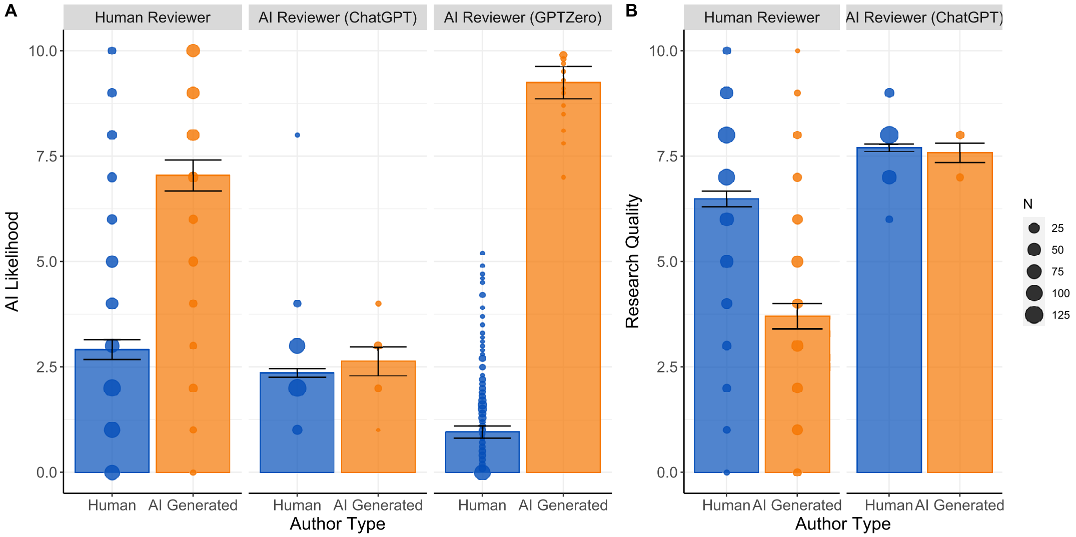

Figure 1. *Research likelihood to be generated by AI (panel A) and research quality (panel B) of abstracts written by human or AI authors as evaluated by human or AI reviewers. Barplots represent mean scores in each group, and error bars represent 95% confidence intervals. Dots represent individual observations, and the size of the dots illustrates the number of observations.*

distinct categories, as well as the reviewer (given that human reviewers evaluated several abstracts) were treated as random effects to account for the potentially interrelated error variance.

As shown in Figure 1A, the mean rating for human-generated abstracts to be generated by AI on a scale from 0 (‘very unlikely’) to 10 (‘very likely’) was $M = 2.73$ ($Md = 2, SD = 2.19$), which was significantly lower when compared to AI-generated abstracts ($M = 6.69, Md = 8, SD = 2.92$), Mann–Whitney U test: $U = 26,216, z = 15.96, p < 0.001$. However, as indicated by the significant interaction effect of author type and reviewer type in Models 1 and 2 (see Table 1), this difference is entirely driven by human reviewers (simple slope: unstandardized $B = 4.15, SE = 0.67, p < 0.001$), supporting H1. In contrast, the AI reviewer does not discriminate between human- and AI-generated abstracts (simple slope: $B = 0.28, SE = 0.49, p = 0.563$). This rejects H2 that the AI reviewer would be more likely to identify AI-generated abstracts as AI-generated than human reviewers; actually, the data suggest that the opposite is true. Regarding the quality of the research, rated from 0 (‘very poor quality’) to 10 (‘very high quality’), research made-up and reported by AI ($M = 4.01, Md = 4, SD = 2.42$) was rated as significantly lower in quality when compared to actual research reported in human-written abstracts ($M = 6.88, Md = 7, SD = 1.82$), Mann–Whitney U test: $U = 29,865, z = 15.00, p < 0.001$. Although this finding is in support of H3, the mixed effects regression analyses reported in Models 3 and 4 (Table 1) again show that this difference is qualified by a significant interaction effect between author type and reviewer type (for a visualization, see Figure 1B). Specifically, human reviewers evaluated the research reported in human-written abstracts more positively than the research reported in AI-generated abstracts (simple slope: $B = 2.79, SE = 0.13, p < 0.001$). In contrast, the AI reviewer did not discriminate in the research quality reported in human- versus AI-generated abstracts (simple slope: $B = 0.12, SE = 0.38, p = 0.752$). This finding is in support of H4.

3.2. *Exploratory analyses*

GPTZero. The results indicate that ChatGPT is poor at discriminating between human- and AI-generated abstracts. However, there are commercially available LLM detector tools, such as GPTZero (Tian and Cui, 2023), that have been designed to identify the extent to which texts were written with the

https://doi.org/10.1017/jdm.2024.24 Published online by Cambridge University Press

Table 1. *Regression models predicting the perceived likelihood of an abstract to be AI-generated (Models 1 and 2) and the perceived quality of an abstract (Models 3 and 4) using human and ChatGPT ratings.*

	AI likelihood		Quality	
	Model 1	Model 2	Model 3	Model 4
(Intercept)	2.346	2.354	7.699***	7.739***
	(1.241)	(1.213)	(1.129)	(1.138)
Author type:	0.283	0.274	−0.120	−0.161
AI-generated	(0.488)	(0.494)	(0.380)	(0.385)
Reviewer type:	0.519	0.499	−1.212	−1.198
Human reviewer	(1.246)	(1.214)	(1.134)	(1.141)
Author type ×	3.871***	3.896***	−2.672***	−2.686***
Reviewer type	(0.516)	(0.515)	(0.402)	(0.401)
Author seniority:		0.401*	−0.241	
Junior author		(0.169)		(0.132)
Author gender:		−0.485**		0.144
Female		(0.176)		(0.137)
Author gender:		−1.058		1.924
Diverse		(1.517)	(1.187)	
Author gender:		0.469		−0.236
Prefer not to say		(0539)		(0.418)
Random effects				
σ^2	4.15	4.13	2.53	2.52
τ00	1.52$_{Reviewer}$	1.44$_{Reviewer}$	1.26$_{Reviewer}$	1.28$_{Reviewer}$
	0.07$_{Research\ topic}$	0.07$_{Research\ topic}$		
ICC	0.28	0.27	0.33	0.34
N	218$_{Reviewer}$	218$_{Reviewer}$	218$_{Reviewer}$	218$_{Reviewer}$
	22$_{Research\ topic}$	22$_{Research\ topic}$		
Observations	937	937	937	937
Marginal R^2/Conditional R^2	0.354/0.569	0.357/0.573	0.371/0.545	0.380/0.546

Note: We excluded the random effect of the research topic from Models 1 and 2, as they did not contribute to explaining any variance in the outcome.
*$p < 0.05$,
**$p < 0.01$,
***$p < 0.001$.

assistance of LLM. To explore whether GPTZero is indeed better at detecting AI-generated abstracts, each abstract was fed into GPTZero one by one, and each received a score between 0 and 100, indicating the likelihood that the text is entirely AI-generated.

To test H2 using GPTZero as the AI reviewer, we conducted the same mixed effects regressions with the perceived AI likelihood as the outcome as reported above. Prior to the analysis, we rescaled the GPTZero score from 0–100 to 0–10, to make it comparable to the ratings by human reviewers. As shown in Table 2, we find a main effect of author type, indicating that both human and AI reviewers were able to accurately distinguish between human and AI-generated abstracts. In fact, the significant interaction between author type and reviewer type suggests that GPTZero was even better than human reviewers in discerning authorship (see Figure 1A). Hence, when using GPTZero instead of ChatGPT, we find support for H2.

Rating agreement. Human reviewers often show variability in their assessments, indicating a lack of agreement in the evaluation of scientific work. Exploring the agreement among human

Table 2. *Regression models predicting the perceived likelihood of an abstract to be AI-generated using human and GPTZero ratings.*

	AI likelihood	
	Model 1	Model 2
(Intercept)	0.949	0.894
	(1.264)	(1.164)
Author type: AI-generated	8.293***	8.343***
	(0.498)	(0.506)
Reviewer type: Human reviewer	1.935	1.900
	(1.271)	(1.165)
Author type × Reviewer type	−4.129***	−4.114***
	(0.527)	(0.528)
Author seniority: Junior author		0.517**
		(0.173)
Author gender: Female		−0.453*
		(0.180)
Author gender: Diverse		−0.404
		(1.551)
Author gender: Prefer not to say		0.461
		(0.551)
Random effects		
σ^2	4.35	4.33
τ00	1.58 *Research*	1.32 *Research* 0.10 *Research topic*
ICC	0.27	0.25
N	218 *Research*	218 *Research* 22 *Research topic*
Observations	937	937
Marginal R^2/conditional R^2	0.475/0.615	0.485/0.612

*Note:***p* < 0.05,
***p* < 0.01,
****p* < 0.001.

reviewers when evaluating the same abstract, we found the quality scores showed a medium-sized positive correlation ($r = 0.38, 95\% CI[0.26, 0.48], t(229) = 6.14, p < 0.001$).[5] Similarly, the AI likelihood scores from 2 human reviewers exhibited a medium-sized positive correlation ($r = 0.33, 95\% CI[0.21, 0.44], t(229) = 5.23, p < 0.001$). Additionally, using Cohen's Kappa for human reviewer scores, we found a 20.8% agreement in quality assessments (*N* Abstracts = 231) and a 16% agreement in AI likelihood assessments. So, indeed, agreement between human reviewers was rather low.

Investigating the association between AI and human reviewers' assessments, we observed a positive small-to-medium correlation in the quality scores assigned by both AI and human reviewers ($r = 0.23, 95\% CI[0.11, 0.35], t(243) = 3.68, p < 0.001$). The AI likelihood scores from AI and 1 randomly selected human reviewer exhibited a positive but very small and statistically insignificant correlation ($r = 0.08, 95\% CI[-0.04, 0.21], t(243) = 1.29, p = 0.200$), which corresponds to the

[5] We conducted the same analyses among junior reviewers ($r = 0.43, 95\% CI[0.23, 0.59], t(80) = 4.22, p < 0.001$) and senior reviewers ($r = 0.46, 95\% CI[0.31, 0.59], t(126) = 5.84, p < 0.001$), and we found similar results.

https://doi.org/10.1017/jdm.2024.24 Published online by Cambridge University Press

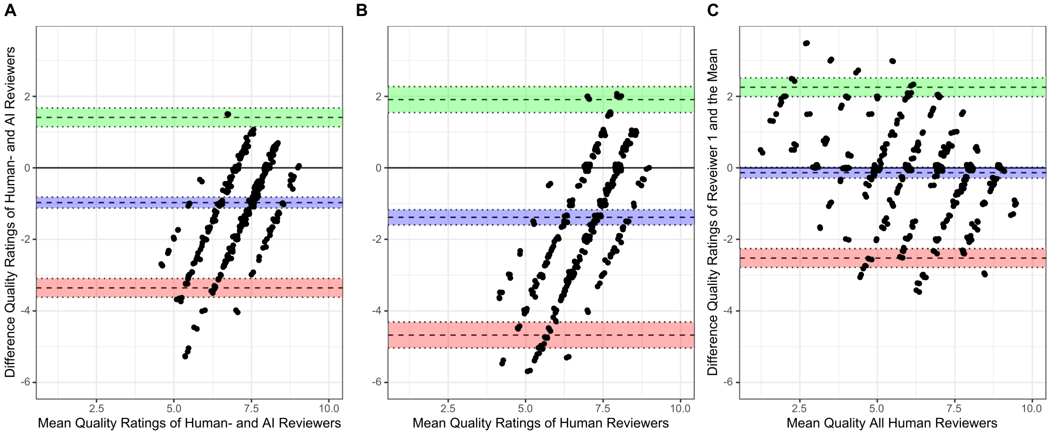

Figure 2. *Bland–Altman plots of agreement in research quality ratings between AI and the mean rating of human and AI reviewers (A), between AI and the mean rating of several human reviewers (B), and between 2 randomly paired human reviewers (C).*

findings reported in Table 1 (Models 1 and 2). Furthermore, utilizing Cohen's Kappa for AI reviewer and 1 randomly selected human reviewer, we identified a 22% agreement in quality assessments and a 17.6% agreement in AI likelihood assessments for the same set of abstracts and rater. Notably, these figures are very similar to the agreement between the 2 human reviewers reported above. Taken together, the results indicate that reviewer agreement is low but comparable between human–human and human–AI reviewer pairs.

To further explore whether LLMs may be useful for screening conference abstracts, and if so, whether there is agreement with human reviewers in the classification of very good or very bad abstracts, we utilized Bland–Altman plots (Bland and Altman, 1986) to visually assess the agreement between human and AI reviewers for different quality levels. Specifically, these plots depict the differences between the 2 methods of measurement in comparison to their average evaluation. Figure 2 illustrates that the AI quality ratings are biased, and the mean human-perceived quality is lower than AI ratings, see panels A and B (AI vs. human reviewers) in comparison to panel C (only human reviewers). Nevertheless, as the differences between human and AI quality ratings become smaller with larger (average) quality ratings in panels A and B, current LLMs may be more reliable in prescreening high-quality research abstracts than low-quality research abstracts.

4. Discussion

Our study contributes to the growing discussion on the utilization of LLMs for improving the taxed academic review process (e.g., Liang et al., 2024; Liu and Shah, 2023). We focused on 2 main aspects of the review process: authorship detection and quality evaluation.

First, our results corroborate previous findings on probabilistic detection of AI-generated abstracts, suggesting that researchers are good, and in our case even better than ChatGPT, at differentiating human- versus AI-generated content. However, our exploratory analysis revealed that AI classifier tools like GPTZero can accurately detect AI-generated content, even exceeding human reviewers' discriminatory ability.

Second, we find that human reviewers evaluate the real abstracts by human authors more positively than the made-up abstracts by AI authors, whereas the AI reviewer failed to differentiate between them. While the evaluations of human and AI reviewers are positively but only weakly related, the same is

https://doi.org/10.1017/jdm.2024.24 Published online by Cambridge University Press

true for the evaluations of different human reviewers. ChatGPT offers inflated quality evaluations of both original and generated work, failing to distinguish high- and low-quality output. The findings further suggest that human and AI reviewers are more aligned when it comes to evaluating the very best abstracts than the low-quality abstracts.

4.1. Implications

The results suggest that at this point LLMs cannot substitute human reviewers in the peer review process. Nevertheless, there might be opportunities for human–AI collaboration. For instance, there may be an increasing need to differentiate whether certain texts were generated by humans or AI. Our results suggest that AI-generated scientific abstracts can well be identified by GPTZero. Such specialized software may serve journals or editors as a screening device, but it is important to bear in mind that its accuracy is highly dependent on text length and training data (OpenAI et al., 2023). While it is true that the review process is time-consuming and that some cases take more time to review than others, utilizing LLMs for preprocessing and initial screening can provide valuable insights across all types of submissions. Although it is unlikely for an extended paper to be purely AI-generated, such scenarios are becoming more common. Many conferences only require an abstract for submission, and the sheer volume of submissions makes high-quality and fast reviews more challenging. Utilizing LLMs to reduce the burden of filtering and preprocessing submissions can significantly improve the overall review quality by freeing up more time and resources for reviewers to focus on in-depth evaluations for the remaining submissions. Future research should investigate the detection accuracy in the case of full-length scientific papers, with only parts being written by LLMs and in different scientific disciplines.

In the context of quality evaluations, human researchers were much better at indicating gradation of content quality and clearly ranked AI-generated content as much lower quality than human-written text. As human reviewers knew that 1 abstract was AI-generated, it could be that such lower quality ratings result from the human reviewers' belief that those abstracts were generated by AI, reflecting a general competence bias against AI-generated content (e.g., Böhm et al., 2023). Our results should be interpreted with caution when generalizing to settings in which reviewers are not aware that content could be generated by AI (which may be less affected by AI aversion).

Moreover, we confirm the tendency of AI-assisted evaluations to present inflated quality ratings for both human-generated and AI-assisted abstracts, significantly favoring its own output. In the work of Latona et al. (2024), this had a positive effect on the paper acceptance rate, whereby a submission that received an AI-assisted review was 4.9 percentage points more likely to be accepted. Combined, our results stress the importance of examining the impact that LLMs have on the review process and introducing guidelines for the integration of LLMs.

Importantly, in the present study, the AI reviewers had good agreement with human reviewers when it came to identifying high-quality research abstracts. As such, AI could support the review process as a prescreening tool, before human reviewers engage in more fine-grained evaluation (e.g., for selecting conference presentations). Agreement between human and AI reviewers could be further increased when providing more specific and less subjective evaluation criteria (e.g., Liu and Shah, 2023). In the present study, reviewers were simply asked to evaluate the 'quality' of the presented research, which is quite broad and subjective.

AI-based support in scientific reviewing has the potential to reduce or even eliminate well-known biases in the reviewing system, such as evaluations based on author identifiers, null results, or certain buzzwords (Liu and Shah, 2023). However, this potential should be viewed cautiously and critically, given the documented instances of AI models inheriting and even amplifying biases from their training data (Bernhardt et al., 2022; Hartmann et al., 2023; Scherrer et al., 2024). Future research should investigate under what conditions and with what prompts about evaluation criteria such biased reviews are likely to disappear when evaluated by AI reviewers.

4.2. Limitations

There are several limitations to our study. Our sample of participants suffers from self-selection bias since the results are based on responses from conference participants who consented to review AI-generated work.[6] While the call for participation was open to all conference participants, it is quite likely that participants who opted into the study were already familiar with and interested in LLMs. Such reviewers may have been particularly knowledgeable in distinguishing between actual human-generated abstracts and made-up AI-generated abstracts. Furthermore, we only compared purely AI-generated and human-written abstracts. Our primary intention was to isolate the effects of LLMs in the abstract generation and review processes, thereby limiting the potential impact of human intervention. Researchers commonly employ LLMs to improve many parts of their writing, such as language refinement, clarity, and idea development (Liang et al., 2024a,b). We did not have data on the extent or method by which researchers used LLMs to improve their submissions. Therefore, our study cannot account for the complex ways AI may have contributed to the quality of human-written abstracts. However, we did evaluate the extent to which authors have potentially used ChatGPT using GPTZero, and the overall prediction was rather low (8.9%, see Model 2 in Table 2). Moreover, we have used the same prompt for both human reviewers and ChatGPT for evaluation. As described above, some prompts were quite abstract and lacked specific evaluation criteria, which could have been an issue for ChatGPT evaluations. A precise and unambiguous prompt may better demonstrate the limits and capabilities of LLMs, and some researchers (e.g., Liang et al., 2024; Liu and Shah, 2023) have either given very specific criteria or dynamically changed the prompt to probe for the most accurate answer. Finally, our results are naturally limited to the current technology (i.e., GPT-3.5 and GPT-4 as used in the present study) and knowledge of the sample of reviewers.[7] Improved LLMs, some of which are fine-tuned on scientific data (e.g., SciBERT Beltagy et al., 2019), may soon overcome some of the weaknesses of AI-based reviews.

5. Conclusion

LLMs will undoubtedly impact the scientific peer review process, and our results suggest that the impact is likely to be positive and useful for the scientific reviewing system. While we should not expect to outsource scientific review to LLMs such as ChatGPT in the nearest future, we already know which tasks can be delegated. Prescreening papers can identify and elevate highly robust scientific work, bypassing known reviewer biases. In sum, at its current state, LLMs have a lot to offer to the scientific review process, both in terms of publishing efficiency and diversity of the author pool.

Data availability statement. The ethics approval, study materials, data, and analysis code are publicly accessible in online (https://osf.io/nje23).

Author contributions. A.S. and H.H. have contributed equally to this work.

Competing interests. The authors declare no competing interests in this work.

[6] We compared the gender and seniority distribution among conference participants and those who consented to be part of the study. Regarding gender, 38% of the overall conference participants were female, similar to the 37% of those who consented to be part of the study. Additionally, 69% of abstracts were submitted by individuals with a Ph.D., while 62% of the consented participants hold a Ph.D.

[7] To stress-test our results and explore whether newer versions of LLMs produce different outcomes, we have collected another round of data using GPT-4o. Our findings indicate no significant main or interaction effects between author type and the version of GPT models—newer versus older—in terms of detecting AI-assisted content (main effect: $B = 0.362, SE = 2.008, p = 0.857$, interaction: $B = 0.427, SE = 0.599, p = 0.476$) and judging abstract quality (main effect: $B = -0.292, SE = 1.602, p = 0.856$, interaction: $B = -0.708, SE = 0.478, p = 0.139$).

https://doi.org/10.1017/jdm.2024.24 Published online by Cambridge University Press

References

Aczel, B., Szaszi, B., & Holcombe, A. O. (2021). A billion-dollar donation: Estimating the cost of researchers' time spent on peer review. *Research Integrity and Peer Review*, *6*(1), 14. https://doi.org/10.1186/s41073-021-00118-2

Beltagy, I., Lo, K., & Cohan, A. (2019). Scibert: Pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 3615–3620). Hong Kong: Association for Computational Linguistics.

Bernhardt, M., Jones, C., & Glocker, B. (2022). Potential sources of dataset bias complicate investigation of underdiagnosis by machine learning algorithms. *Nature Medicine*, *28*(6), 1157–1158. https://doi.org/10.1038/s41591-022-01846-8.

Böhm, R., Jörling, M., Reiter, L., & Fuchs, C. (2023). People devalue generative AI's competence but not its advice in addressing societal and personal challenges. *Communications Psychology*, *1*(1), 1–10. https://doi.org/10.1038/s44271-023-00032-x

D'Andrea, R., & O'Dwyer, J. P. (2017). Can editors save peer review from peer reviewers?. *PLOS ONE*, *12*(10), e0186111. https://doi.org/10.1371/journal.pone.0186111

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., Baabdullah, A. M., Koohang, A., Raghavan, V., Ahuja, M., Albanna, H., Albashrawi, M. A., Al-Busaidi, A. S., Balakrishnan, J., Barlette, Y., Basu, S., Bose, I., Brooks, L., Buhalis, D., . . . Wright, R. (2023). Opinion paper: "so what if ChatGPT wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational AI for research, practice and policy. *International Journal of Information Management*, *71*, 102642. https://doi.org/10.1016/j.ijinfomgt.2023.102642

Foltýnek, T., Meuschke, N., & Gipp, B. (2019). Academic plagiarism detection: A systematic literature review. *ACM Computing Surveys*, *52*(6), Article no. 112, 1–42. https://doi.org/10.1145/3345317

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2023). Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. *NPJ Digital Medicine*, *6*(1), 1–5. https://doi.org/10.1038/s41746-023-00819-6

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., Yue, J., & Wu, Y. (2023). How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection. https://doi.org/10.48550/arXiv.2301.07597

Haffar, S., Bazerbachi, F., & Murad, M. H. (2019). Peer review bias: A critical review. *Mayo Clinic Proceedings*, *94*(4), 670–676. https://doi.org/10.1016/j.mayocp.2018.09.004

Hartmann, J., Schwenzow, J., & Witte, M. (2023). The political ideology of conversational AI: Converging evidence on ChatGPT's proenvironmental, left-libertarian orientation. https://doi.org/10.48550/ARXIV.2301.01768

Hosseini, M., & Horbach, S. P. J. M. (2023). Fighting reviewer fatigue or amplifying bias? Considerations and recommendations for use of ChatGPT and other large language models in scholarly peer review. *Research Integrity and Peer Review*, *8*(1), 4. https://doi.org/10.1186/s41073-023-00133-5

Huisman, J., & Smits, J. (2017). Duration and quality of the peer review process: The author's perspective. *Scientometrics*, *113*(1), 633–650. https://doi.org/10.1007/s11192-017-2310-5

Kovanis, M., Porcher, R., Ravaud, P., & Trinquart, L. (2016). The global burden of journal peer review in the biomedical literature: Strong imbalance in the collective enterprise. *PLOS ONE*, *11*(11), e0166387. https://doi.org/10.1371/journal.pone.0166387

Latona, G. R., Ribeiro, M. H., Davidson, T. R., Veselovsky, V., & West, R. (2024). The AI review lottery: Widespread AI-assisted peer reviews boost paper scores and acceptance rates. https://doi.org/10.48550/arXiv.2405.02150

Laycock, H., & Bailey, C. R. (2019). The influence of first author sex on acceptance rates of submissions to anaesthesia cases. *Anaesthesia*, *74*(11), 1432–1438. https://doi.org/10.1111/anae.14797

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology*, *64*(1), 2–17. https://doi.org/10.1002/asi.22784

Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., Chen, L., Ye, H., Liu, S., Huang, Z., McFarland, D. A., & Zou, J. Y. (2024a). Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. https://doi.org/10.48550/arXiv.2403.07183

Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D. Y., Yang, X., Vodrahalli, K., He, S., Smith, D. S., Yin, Y., McFarland, D. A., & Zou, J. (2024). Can large language models provide useful feedback on research papers? A large-scale empirical analysis. NEJM AI, *1*(8). https://doi.org/10.1056/aioa2400196

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., & Zou, J. Y. (2024b). Mapping the increasing use of LLMs in scientific papers. http://arxiv.org/abs/2404.01268

Lindsay, G. W. (2023). LLMs are not ready for editorial work. *Nature Human Behaviour*, *7*(11), 1814–1815. https://doi.org/10.1038/s41562-023-01730-6

Liu, R., & Shah, N. B. (2023). ReviewerGPT? An exploratory study on using large language models for paper reviewing. https://doi.org/10.48550/arXiv.2306.00622

Ma, Y., Liu, J., Yi, F., Cheng, Q., Huang, Y., Lu, W., & Liu, X. (2023). AI vs. human – differentiation analysis of scientific content generation. https://doi.org/10.48550/arXiv.2301.10416

Murray, D., Siler, K., Larivière, V., Chan, W. M., Collings, A. M., Raymond, J., & Sugimoto, C. R. (2019). Author-reviewer homophily in peer review. https://doi.org/10.1101/400515

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., . . . Zoph, B. (2023). GPT-4 Technical Report (arXiv:2303.08774; Version 4). https://doi.org/10.48550/arXiv.2303.08774

https://doi.org/10.1017/jdm.2024.24 Published online by Cambridge University Press

Petrescu, M., & Krishen, A. S. (2022). The evolving crisis of the peer-review process. *Journal of Marketing Analytics*, *10*(3), 185–186. https://doi.org/10.1057/s41270-022-00176-5

Sabel, B. A., Knaack, E., Gigerenzer, G., & Bilc, M. (2023). Fake publications in biomedical science: Red-flagging method indicates mass production. https://doi.org/10.1101/2023.05.06.23289563

Scherrer, N., Shi, C., Feder, A., & Blei, D. (2024). Evaluating the moral beliefs encoded in llms. *Advances in Neural Information Processing Systems*, *36*, 51778–51809.

Schulz, R., Barnett, A., Bernard, R., Brown, N. J. L., Byrne, J. A., Eckmann, P., Gazda, M. A., Kilicoglu, H., Prager, E. M., Salholz-Hillel, M., ter Riet, G., Vines, T., Vorland, C. J., Zhuang, H., Bandrowski, A., & Weissgerber, T. L. (2022). Is the future of peer review automated? *BMC Research Notes*, *15*(1), 203. https://doi.org/10.1186/s13104-022-06080-6

Smith, O. M., Davis, K. L., Pizza, R. B., Waterman, R., Dobson, K. C., Foster, B., Jarvey, J. C., Jones, L. N., Leuenberger, W., Nourn, N., Conway, E. E., Fiser, C. M., Hansen, Z. A., Hristova, A., Mack, C., Saunders, A. N., Utley, O. J., Young, M. L., & Davis, C. L. (2023). Peer review perpetuates barriers for historically excluded groups. *Nature Ecology & Evolution*, *7*(4), 512–523. https://doi.org/10.1038/s41559-023-01999-w

Tian, E., & Cui, A. (2023). Gptzero: Towards detection of AI-generated text using zero-shot and supervised methods. https://gptzero.me

Wicherts, J. M. (2016). Peer review quality and transparency of the peer-review process in open access and subscription journals. *PLOS ONE*, *11*(1), e0147913. https://doi.org/10.1371/journal.pone.0147913

Cite this article: Shcherbiak, A., Habibnia, H., Böhm, R., and Fiedler, S. (2024). Evaluating science: A comparison of human and AI reviewers. *Judgment and Decision Making*, e21. https://doi.org/10.1017/jdm.2024.24