Neuroscience

ChatGPT identifies gender disparities in scientific peer review

Jeroen P. H. Verharen

Department of Molecular and Cell Biology and Helen Wills Neuroscience Institute, University of California, Berkeley, USA

Reviewed Preprint

Revised by authors after peer review.

About eLife's process

Reviewed preprint version 2
October 4, 2023 (this version)

Reviewed preprint version 1
August 8, 2023

Posted to bioRxiv
July 19, 2023

Sent for peer review
June 22, 2023

Abstract

The peer review process is a critical step in ensuring the quality of scientific research. However, its subjectivity has raised concerns. To investigate this issue, I examined over 500 publicly available peer review reports from 200 published neuroscience papers in 2022-2023. OpenAI's generative artificial intelligence *ChatGPT* was used to analyze language use in these reports. It demonstrated superior performance compared to traditional lexicon- and rule-based language models. As expected, most reviews for these published papers were seen as favorable by *ChatGPT* (89.8% of reviews), and language use was mostly polite (99.8% of reviews). However, this analysis also demonstrated high levels of variability in how each reviewer scored the same paper, indicating the presence of subjectivity in the peer review process. The results further revealed that female first authors received less polite reviews than their male peers, indicating a gender bias in reviewing. In addition, published papers with a female senior author received more favorable reviews than papers with a male senior author, for which I discuss potential causes. Together, this study highlights the potential of generative artificial intelligence in performing natural language processing of specialized scientific texts. As a proof of concept, I show that *ChatGPT* can identify areas of concern in scientific peer review, underscoring the importance of transparent peer review in studying equitability in scientific publishing.

eLife assessment

This study used ChatGPT to assess certain linguistic characteristics (sentiment and politeness) of 500 peer reviews for 200 neuroscience papers published in Nature Communications. The vast majority of reviews were polite, but papers with female first authors received less polite reviews than papers with male first authors, whereas papers with a female senior author received more favourable reviews than papers with a male senior author. Overall, the study is an **important** contribution to work on gender bias, and the evidence for the potential utility of generative AI programs like ChatGPT in meta-research is **solid**.

Introduction

The peer review process is a crucial step in the publication of scientific research, where manuscripts are evaluated by independent experts in the field before being accepted for publication. This process helps ensure the quality and validity of scientific research and is a cornerstone of scientific integrity. Despite its importance, concerns have been raised regarding subjectivity in this process that may affect the fairness and accuracy of evaluations[1-5]. Indeed, most journals engage in single-blind peer review, in which the reviewers have information about the authors of the paper, but not vice versa. While some studies have found evidence of disparities in peer review as a result of gender bias, the scope and methodology of these studies are often limited[6-7]. One larger study, performed by an ecology journal, found no evidence of gender bias in reviewing, but did find a bias against non-English-speaking first authors[8]. Additionally, other factors, such as the seniority and institutional affiliation of authors, may influence the evaluation process and lead to biased assessments of research quality[6]. As such, papers from more prestigious research institutions may receive better reviews[9]. It is crucial to identify potential sources of disparity in the reviewing process to maintain scientific integrity and find areas for improvement within the scientific pipeline.

Natural language processing tools have shown promise in analyzing large amounts of textual data and extracting meaningful insights from evaluations[10-12]. However, applying these tools to scientific peer review has been challenging due to the specialized construction and language use in such reports. A recent study that manually annotated language use in peer reviews has shown great potential[13], but algorithms struggled to perform well in this task[14-15]. Recent advances in generative artificial intelligence, such as OpenAI's *ChatGPT*, offer new possibilities for studying scientific peer review. These models can process vast amounts of text and provide accurate sentiment scores and language use metrics for individual sentences and documents. As such, using generative artificial intelligence to study scientific peer review may ultimately help improve the overall quality and fairness of scientific publications and identify areas of concern in the way towards equitable academic research.

This study had three main objectives. The first aim was to test whether the latest advances in generative artificial intelligence, such as OpenAI's *ChatGPT*, can be used to analyze language use in specialized scientific texts, such as peer reviews. The second aim was to explore subjectivity in peer review by looking at consistency in favorability across reviews for the same paper. The last aim was to test whether the identity of the authors, such as institutional affiliation and gender, affect the favorability and language use of the reviews they receive.

Results

An analysis of scientific peer review

Nature Communications has engaged in transparent peer review since 2016, giving authors the option to (and since 2022 requiring authors to) publish the peer review history of their paper[16]. To explore language use in these reports, I downloaded the primary (i.e., first-round) reviews from the last 200 papers in the neuroscience field published in this journal. This yielded a total of 572 reviews from 200 papers, with publications dates ranging from August 2022 to February 2023. Additional metrics of these papers were manually collected (**Fig. 1a** and **1b**), including the total time until paper acceptance, the subfield of neuroscience, the geographical location and QS World Ranking score of the senior author's institutional affiliation, the gender of the senior author, and whether the first author had a male or female name (see **Methods** for more

information on classifications and a rationale for the chosen metrics). These metrics were collected to test whether they influenced the favorability and language use of the reviews that a paper received.

Sentiment analysis

To assess the sentiment and language use of each of the peer review reports, I asked OpenAI's generative artificial intelligence *ChatGPT* to extract two scores from each of the reviews (**Fig. 2a**). The first score was the *sentiment score*, and measures how favorable the review is. This metric ranges from -100 (negative) to 0 (neutral) to +100 (positive). Sentiment reflects the reviewer's opinion about the paper and is what presumably drives the decision for a paper to be accepted or rejected. The second score was the *politeness score*, which evaluates how polite a review's language is, measured on a scale from -100 (rude) to 0 (neutral) to +100 (polite). *ChatGPT* was able to extract sentiment and politeness scores for all of the 572 reviews, and usually included a reasoning of how it established the score (**Supplementary Fig. 1**).

The accuracy and consistency of the generated scores was validated in four different ways. First, for a representative sample of the reviews, I read both the review and *ChatGPT*'s reasoning of how it came to the scores (for examples see **Supplementary Fig. 1**). I established that the algorithm was able to extract the most important sentences from each of the reviews and to provide a plausible score. Second, since generative artificial intelligence can provide different answers every time it is prompted, the algorithm was asked to provide scores for each review twice. This yielded a significant correlation between the first and second iteration of scoring ($P < 0.0001$ for both sentiment and politeness scores; **Supplementary Fig. 2**); the average of the two scores was used for all subsequent analyses in this paper. Third, manipulated reviews (in which I manually re-wrote a 'neutral' review in a more rude, polite, negative or positive manner) were input into *ChatGPT*, which confirmed that this changed the review's politeness and sentiment scores, respectively (**Supplementary Fig. 3**). Finally, for a subset of reviews, *ChatGPT*'s scores were compared to that of seven human scorers that were blinded to the algorithm's scores (**Supplementary Fig. 4**). Interestingly, there was high variability across human scorers, but their average score had a high correlation to that of *ChatGPT* (linear regression for sentiment score: $R^2 = 0.91$, $P = 0.0010$; for politeness score: $R^2 = 0.70$, $P = 0.018$). Importantly, *ChatGPT* was superior to the lexicon- and rule-based algorithms *TextBlob*[17] and *VADER*[18] in scoring a review's sentiment; both these algorithms did not not significantly predict the average human-scored sentiment (*TextBlob*: $R^2 = 0.13$, $P = 0.42$; *VADER*: $R^2 = 0.07$, $P = 0.56$). Together, these validations indicate that *ChatGPT* can accurately score the sentiment and politeness of scientific peer reviews, and does so better than other available tools.

The majority of the 572 peer reviews (89.9%) were of positive sentiment; 7.9% were negative; 2.3% were neutral (i.e., a sentiment score of 0) (**Fig. 2b**). 99.8% of reviews were deemed polite by the algorithm (i.e., a positive politeness score), only 1 review was scored as rude (i.e., a negative politeness score; **Fig. 2c**, bottom left inset). A regression analysis indicated a strong relation between the reviews' sentiment and politeness scores (60% of variance explained in a third-degree polynomial regression) (**Fig. 2c**). Thus, the more positive a review, the more polite the reviewer's language generally is. It is important to note here that the papers included in this analysis were ultimately accepted for publication in *Nature Communications*, which has a low acceptance rate of 7.7%. As a result of this selection, there will be an over-representation of positive scores in this analysis (**Fig. 2c**, bottom right inset).

Consistency across reviewers

If a research paper meets certain objective standards of quality, one can reasonably expect that reviewers evaluating that paper would share a common view on its overall sentiment. To investigate if this is the case, I analyzed the consistency across review scores for the same paper

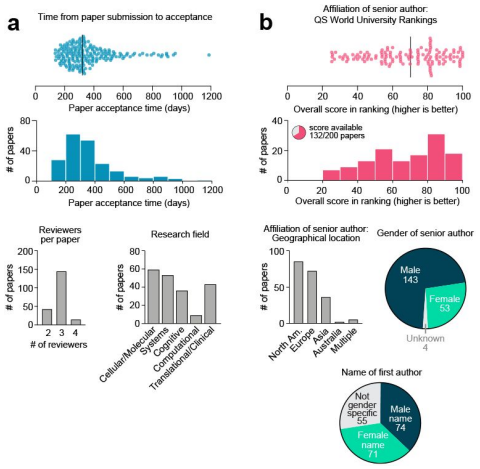


Figure 1

Characteristics of the 200 papers included in this analysis

(a) Paper metrics. (**b**) Author metrics. More information on how these metrics were collected and defined can be found in the **Methods** section.

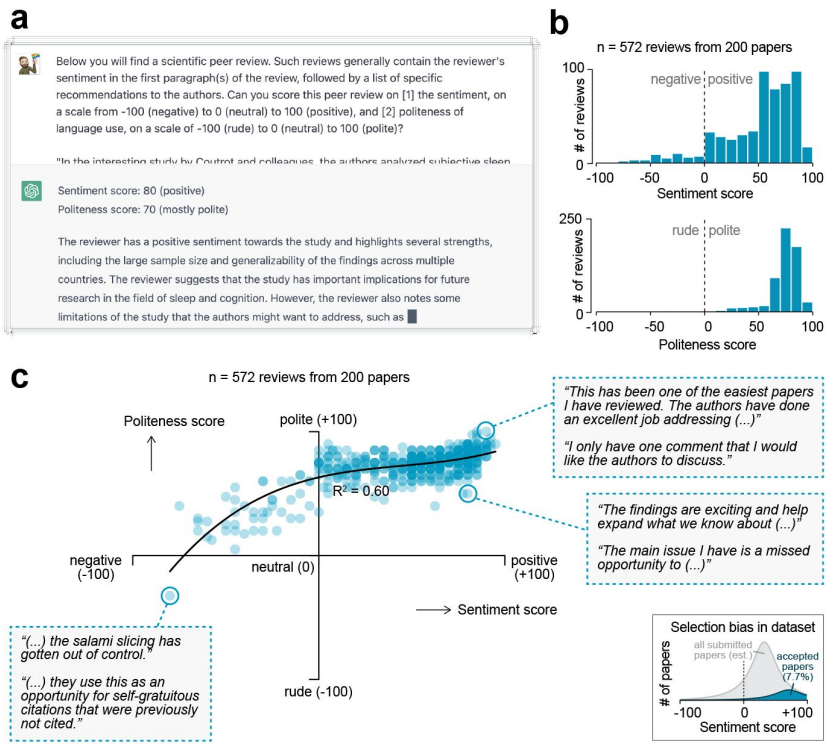


Figure 2

Sentiment analysis on peer review reports using generative artificial intelligence

(a) OpenAI's generative artificial intelligence model *ChatGPT* was used to extract a sentiment and politeness sores for each of the 572 first-round reviews. Shown is an example query and *ChatGPT*'s answer.
(b) Histograms showing the distribution in sentiment (top) and politeness (bottom) scores for all reviews.
(c) Scatter plot showing the relation between sentiment and politeness scores for the reviews (60% variance explained in third-degree polynomial). Insets show excerpts from selected peer reviews. Inset in the bottom right corner is a visual depiction of the expected selection bias in this dataset, as only papers accepted for publication were included in this analysis (gray area represents full pool of published and unpublished papers; not to scale).

(**Fig. 3**). As expected, the overall distribution of sentiment and politeness scores did not differ between the first three reviewers (**Fig. 3a**). Interestingly, a linear regression analysis of sentiment scores across reviewers indicated very low, if any, correlation between the sentiment scores of reviews for the same paper (**Fig. 3b**). The maximum variance explained in sentiment scores between reviewers was 5.5% (between reviewer 1 and 3; the only comparison that reached statistical significance). I also calculated the intra-class correlation coefficient[19] between the different reviewers, which demonstrated poor inter-reviewer reliability of scoring (ICC = 0.055, 95% confidence interval of -0.025 – 0.144). These results indicate high levels of disagreement between the reviewers' favorability of a paper, suggesting that the peer review process is subjective.

I then looked at the relation between a paper's review scores and its acceptance time (i.e., the time from paper submission to acceptance). For this analysis, review scores were first classified as the lowest, median (only for papers with an odd number of reviewers), or highest for a paper (**Fig. 3c**). A linear regression analysis indicated that the median sentiment score was the best predictor of a paper's review time (R^2 = 0.0670, P = 0.0002), followed by the lowest sentiment score (R^2 = 0.1404, P < 0.0001) (**Fig. 3c**, bottom left panels). Interestingly, a paper's highest sentiment score did not significantly predict review time (R^2 = 0.0088, P = 0.1874).

Exploring disparities in peer review

To explore potential sources of disparities in scientific publishing, I correlated the review scores, pooled across all papers, with the different paper and author metrics that were collected earlier (**Fig. 1b**). No significant effects were observed between sentiment and politeness scores across the different subfields of neuroscience (**Fig. 4a**). With respect to the institutional affiliation of the senior author, no effects were observed between the scores and the continent in which the senior author was based (**Fig. 4b**). Additionally, no correlation was observed between the institute's score on the QS World ranking and the paper's sentiment and politeness scores (**Fig. 4c**).

Finally, I looked at how the gender of the first and senior authors may affect a paper's review scores. First authors with a female name received significantly more impolite reviews, but no effect was observed on sentiment (**Fig. 4d**). To study whether these more impolite reviews for female first authors were due to an overall lower politeness score, or due to one or some of the reviewers being more impolite, I split the reviews for each paper by its lowest/ median/highest politeness score. I observed that the lower politeness scores for first authors with a female name was driven by significantly lower low and median scores (**Fig. 4d**, bottom panel). Thus, the least polite reviews a paper received were even more impolite for papers with a female first author. Conversely, female senior authors received significantly higher sentiment scores, indicating more favorable reviews, but these reviews did not differ in terms of politeness (**Fig. 4e**). An analysis of reviews split by lowest/median/highest sentiment score indicated that the reviewer that gave the most favorable review to female senior authors did so with a significantly higher score (**Fig. 4e**, bottom panel). No interactions on scores were observed between the genders of the first and senior authors (**Supplementary Fig. 5**).

Discussion

Peer review is a crucial component of scientific publishing. It helps ensure that research papers are of high quality and have been scrutinized by experts in the field. However, the potential for subjectivity in the peer review process has been an ongoing concern. For example, implicit or explicit bias of reviewers may lead to disparities in peer review scores on the basis of gender or institutional affiliation. In this study, I used natural language processing tools embedded in OpenAI's *ChatGPT* to analyze 572 peer review reports from 200 papers that were accepted for

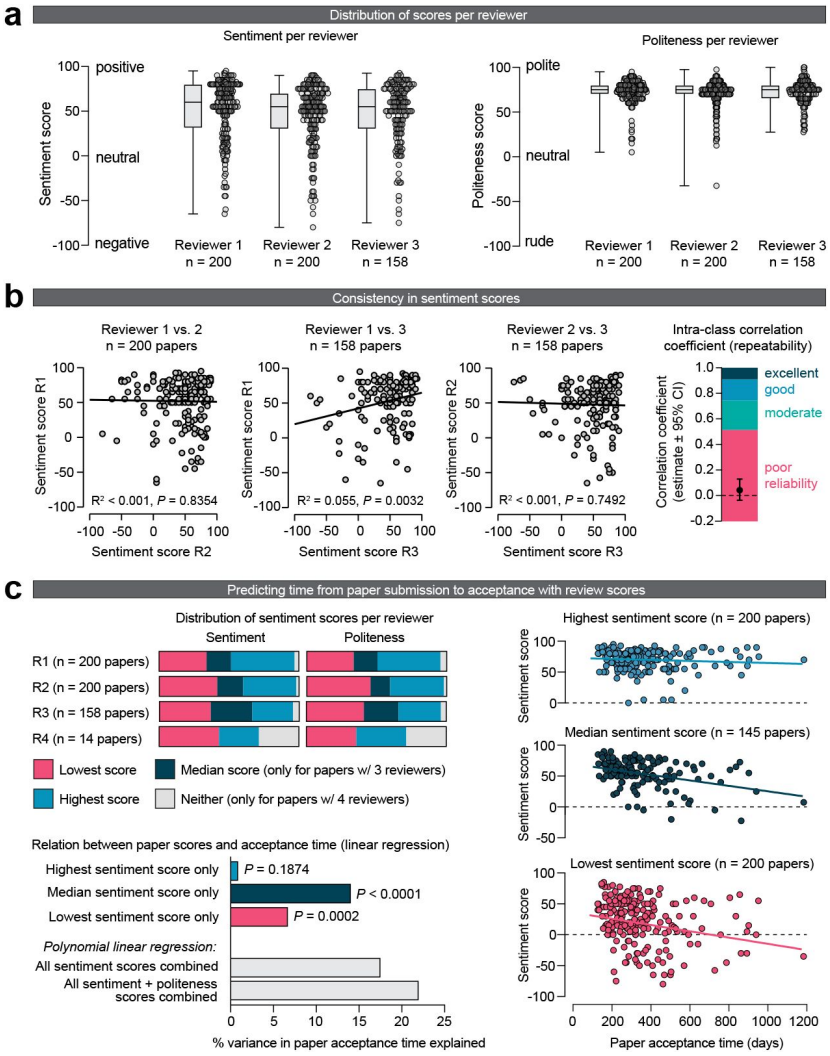


Figure 3

Consistency across reviews

(a) Sentiment (left) and politeness (right) scores for each of the 3 reviewers. The lower sample size for reviewer 3 is because 42 papers received only 2 reviews. No significant effects were observed of reviewer number on sentiment (mixed effects model, $F(1.929, 343.3) = 1.564$, $P = 0.2116$) and politeness scores (mixed effects model, $F(1.862, 331.4) = 1.638$, $P = 0.1977$).
(b) Correlations showing low consistency of sentiment scores across reviews for the same paper. The sentiment scores between reviewers 1 and 3 (middle panel) is the only comparison the reached statistical significance ($P = 0.0032$), albeit with a low amount of variance explained (5.5%). The intra-class correlation coefficient (ICC) measures how similar the review scores are for one paper, without the need to split review up into pairs. An ICC < 0.5 generally indicates poor reliability (i.e., repeatability)[19].
(c) Linear regression indicating the relation between a paper's sentiment scores and the time between paper submission and acceptance. For this analysis, reviews were first split into a paper's lowest, median (only for papers with an odd number of reviews) and highest sentiment score. The lowest and median sentiment score of a paper significantly predicted a paper's review time, but its highest sentiment score did not. Note that the relation between politeness scores and review time were not individually tested given the high correlation between sentiment and politeness, thus having a high chance of finding spurious correlations. The metric '% variance in paper acceptance time explained' denotes the R^2[2] value of the linear regression.

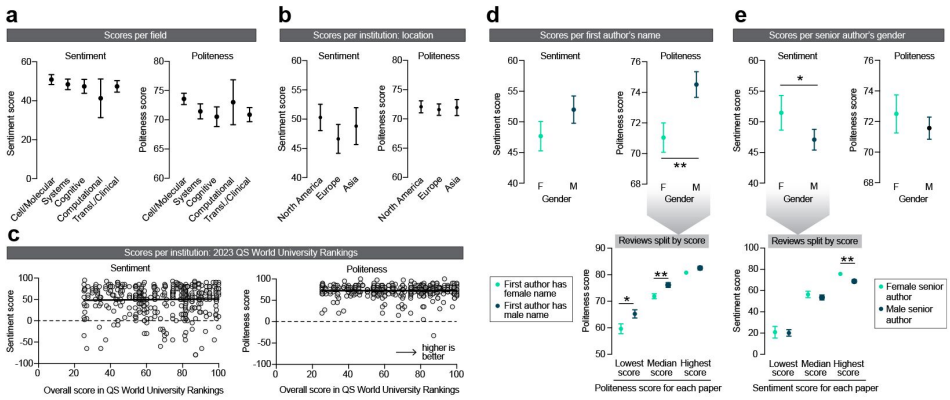


Figure 4

Exploring disparities in peer review

(a) Effects of the subfield of neuroscience on sentiment (left) and politeness (right) scores. No effects were observed on sentiment (Kruskal-Wallis ANOVA, $H = 2.380$, $P = 0.6663$) or politeness (Kruskal-Wallis ANOVA, $H = 8.211$, $P = 0.0842$). $n = 178$, 149, 100, 20, 125 reviews per subfield.
(b) Effects of geographical location of the senior author on sentiment (left) and politeness (right) scores. No effects were observed on sentiment (Kruskal-Wallis ANOVA, $H = 1.856$, $P = 0.3953$) or politeness (Kruskal-Wallis ANOVA, $H = 0.5890$, $P = 0.7449$). $n = 239$, 208, 103 reviews per continent.
(c) Effects of QS World Ranking score of the senior author' institutional affiliation on sentiment (left) and politeness (right) scores. No effects were observed on sentiment (Linear regression, $R^2 = 0.0006$, $P = 0.6351$) or politeness (Linear regression, $R^2 < 0.0001$, $P = 0.9804$). $n = 430$ reviews.
(d) Effects of the first author's name on sentiment (left) and politeness (right) scores. No effects were observed on sentiment (Mann-Whitney test, $U = 19521$, $P = 0.2131$) but first authors with a female name received significantly less polite reviews (Mann-Whitney test, $U = 17862$, $P = 0.0080$, Hodges-Lehmann difference of 2.5). Post-hoc tests on the data split per lowest/median/highest politeness score indicated significantly lower politeness scores for females for the lowest (Mann-Whitney test, $U = 1987$, $P = 0.0103$, Hodges-Lehmann difference of 5) and median (Mann-Whitney test, $U = 1983$, $P = 0.0093$, Hodges-Lehmann difference of 2.5) scores, but not of the highest score (Mann-Whitney test, $U = 2279$, $P = 0.1607$). $n = 206$ (F), 204 (M) reviews for top panels; $n = 71$ (F), 74 (M) papers for lower panel (but $n = 54$ (F), 53 (M) papers for median scores, because not all papers received 3 reviews).
(e) Effects of the senior author's gender on sentiment (left) and politeness (right) scores. Women received more favorable reviews than men (Mann-Whitney test, $U = 28007$, $P = 0.0481$, Hodges-Lehmann difference of 5) but no effects were observed on politeness (Mann-Whitney test, $U = 29722$, $P = 0.3265$). Post-hoc tests on the data split per lowest/median/ highest sentiment score indicated no effect of gender on the lowest (Mann-Whitney test, $U = 3698$, $P = 0.7963$) and median (Mann-Whitney test, $U = 3310$, $P = 0.1739$) sentiment scores, but the highest sentiment score was higher for women (Mann-Whitney test, $U = 2852$, $P = 0.0072$, Hodges-Lehmann difference of 5). $n = 155$ (F), 405 (M) reviews for top panels; $n = 53$ (F), 143 (M) papers for lower panel (but $n = 39$ (F), 102 (M) papers for median scores, because not all papers received 3 reviews). Asterisks indicate statistical significance in Mann-Whitney tests; * $P < 0.05$, ** $P < 0.01$.

publication in *Nature Communications* within the past year. I found that this approach was able to provide consistent and accurate scores, matching that of human scorers. Importantly, *ChatGPT* was superior to the conventional lexicon- and rule-based algorithms *TextBlob* and *VADER* in scoring a review's sentiment.

Such algorithms score a text on the basis of the frequency of certain words, and as such may have trouble analyzing scientific text with specialized constructions and vocabulary[13], as has been shown before[15]. Altogether, the current study serves as a proof of concept for the use of generative artificial intelligence in studying scientific peer review. Such an automated language analysis of peer reviews can be used in different ways, such as after-the-fact analyses (as has been done here), providing writing support for reviewers (for example by implementation in the journal submission portal), or by helping editors pick the best papers or most constructive reviewers.

Notably, there are several limitations to this study. The peer review reports I analyzed are all ultimately accepted for publication in *Nature Communications*, meaning that there is a selection bias in the reviews that were included. As such, papers that have received unfavorable reviews, or papers that have not been sent out for peer review at all, were not included in this analysis. It is unclear what the gender and institutional affiliation distribution is for the papers that were ultimately unpublished. Additionally, this study only focused on the neuroscience field, and the findings may not generalize to other fields. Similarly, it is not clear if the results from this study apply to journals beyond *Nature Communications*. Future studies may expand upon this initial work by incorporating larger sample sizes and encompassing diverse scientific disciplines and journals.

Despite said limitations, this study may reveal several key insights into the peer review process and highlight potential areas of concern within academic publishing. First, this study found that evaluations of the same manuscript varied considerably among different reviewers. This finding suggests that the peer review process may be subjective, with different reviewers having different opinions on the quality and validity of the research. Notably, some level of variability may be expected, for example due to different backgrounds, experiences, and biases of the reviewers. In addition, *ChatGPT* may not always reliably assess a reviews sentiment, adding some spurious inter-reviewer variability. That being said, the extremely low (or even absent) relation between how different reviewers scored the same paper was striking, at least to this author. The inconsistency in the evaluations emphasizes the need for greater standardization in the peer review process, with clear guidelines and protocols that can minimize such discrepancies[20].

I also investigated disparities in peer review based on the institutional affiliation of the senior author of a paper. Specifically, I looked at the geographic location (continent), as well as the score of the institute in the 2023 QS World University Rankings — an imperfect metric of the institute's perceived prestige. This analysis revealed no relation of these two metrics with the sentiment and politeness of the reviews, suggesting that evaluations were not influenced by the geographical location and perceived prestige of the senior author's research institution. This finding is encouraging and suggests that peer review may be based on the quality and merit of the research rather than the authors' research institute. That said, the identity of the peer reviewers is not known, so it cannot be tested whether reviewers have a bias with respect to authors from a more closely related country, culture or institution (i.e., in-group favoritism). In addition, it's important to acknowledge the selection bias present in this study, in which I exclusively considered published papers. This may mask effects resulting from bias with regards to the senior author's institutional affiliation. For example, papers from less prestigious institutions may have a higher rejection rate. To address this concern, future studies could adopt a strategy such as partnering with a journal to analyze the review sentiment associated with both rejected and accepted papers.

This study further found that first authors with a female name received less polite reviews than first authors with a male name, although this did not affect the favorability of their reviews. Regardless, this disparity is worrisome as it may indicate an unconscious gender bias in review writing that may ultimately impact the confidence and motivation of (especially early-stage) female researchers. One may argue that the effect size of gender on politeness scores is small, but given the selection bias in this dataset (**Fig 2c**, bottom right inset), this effect may be larger in the entire pool of reviewed manuscripts (i.e., rejected + accepted). To address this issue, double-blind peer review, where the authors' names are anonymized, could be implemented. Evidence suggests that this is useful in removing certain forms of bias from reviewing[8,9], but has thus far not been widely implemented, perhaps because some studies have cast doubt on its merits[21,22]. Additionally, reviewers could be more mindful of their language use. Indeed, even negative reviews can be written in a polite manner (**Fig. 2c**), and reviewers may want to use *ChatGPT* to extract a politeness score for their review before submitting.

Additionally, female senior authors received more favorable reviews than male senior authors in this pool of accepted papers. This disparity in sentiment score in favor of women may be surprising given the wealth of data showing unconscious bias against women, including in scientific research[23,24]. It is therefore likely that the observed effect is due to selection bias elsewhere in the publishing process. There may be two potential sources of this bias. The first one is that female senior authors may submit better papers to this journal than their male peers, such that the observed gender effect on sentiment is representative for the entire pool of submitted manuscript (i.e., rejected + accepted). This could be the result of institutional barriers that lead to a small, but highly talented pool of female principal investigators[25] that submits better papers than their male peers[26]. Alternatively, women may have a higher level of self-imposed quality control[27], such that men submit more variable quality papers to high-impact journals like *Nature Communications*. In the imperfect process that is editorial decision making, this may lead to the publication of certain lower-quality papers from male senior authors. The second explanation may be related to an (unconscious) selection bias in the editorial process[28], requiring female senior authors to have better papers before being sent out for peer review, or better scores before being invited for a revise-and-resubmit. As such, paper acceptance may serve as a collider variable[29,30], inducing a spurious association between gender of the senior author and sentiment score. Further research is required to investigate the reasons behind this effect and to identify in what level of the publishing system these differences emerge. In **Supplementary Fig. 6**, I propose three different experiments that journals can perform to rule out bias in reviewing or the editorial process.

Together, this study serves as a proof of concept for the use of generative artificial intelligence in analyzing scientific peer review. *ChatGPT* outperformed commonly used natural language processing tools in measuring sentiment of peer reviews, and provides an easy, non-technical way for people to perform language analyses on specialized scientific texts. Using this approach, areas of concern were discovered within the academic publishing system that require immediate attention. One such area is the inconsistency between the reviews of the same paper, indicating some level of subjectivity in the peer review process. Additionally, I uncovered possible gender disparity in academic publishing and reviewing. This research underscores the potential of generative artificial intelligence to evaluate and enhance scientific peer review, which may ultimately lead to a more equitable and just academic system.

Methods

Downloading reviews

Reviewer reports were downloaded from the website of *Nature Communications* in February 2023. Only papers that were categorized under *Biological sciences* > *Neuroscience* were included in this analysis. Not all papers had their primary reviewer reports published; to reach the total of 200 papers with primary review reports, the most recently published 283 papers were considered (published between August 16, 2022 and Feb 17, 2023).

Additional paper metrics were subsequently collected. Paper submission and acceptance date were downloaded from the '*About this article*' section on the paper website. Paper acceptance time was calculated by counting the number of days between these two dates. Research field was manually categorized on the basis of title and abstract of the paper into five different subfields. The affiliation of the senior author was downloaded from the paper website and manually categorized based on continent; if the senior author had affiliations across multiple continents, it was categorized as 'multiple' and not used for further analyses (this was the case for 5 papers). The affiliated institutions' score in the 2023 QS World Ranking was downloaded from the QS World Ranking website (*TopUniversities.com (http://topuniversities.com/)*) in March 2023; the maximum score an institution could receive was 100. Not all institutions were listed in the QS World Ranking, usually because they were not considered an organization of higher education. If a senior author had multiple affiliations, then the affiliation with the highest score was used. Name-based gender categorization of the first author was performed using *ChatGPT* (query: "*Of the following list of international full (first+last) names, can you guess, based on name only, if these people are male, female, or unknown (i.e., name is not gender specific)?*"). As a confirmation, all names that were assigned a gender by *ChatGPT* were verified using the Genderize database (*http://genderize.io*; probability > 0.5). The gender of the senior author was categorized in a similar manner, except that the categorization for gender-unspecific names was manually completed, usually by looking up the senior author on the research institution's website or the author's Google Scholar or Twitter profile. In this manual look up, I tried to find the senior author's preferred pronouns. If not available, I inferred the senior author's gender on the basis of a photograph. I did not find evidence that any of the senior authors included in this analysis identified as non-binary; for 4 senior authors I was not able to find or infer their gender. Note that this gender look-up was performed for the senior author, but not for the first author, for two reasons. First, first authors generally had less of an online presence than seniors authors, and it was challenging to reliably assess their gender identity. Second, I presumed that reviewers are more likely to be familiar with the senior author of papers they review (for example through conferences) than with first authors. As such, reviewers themselves may infer the gender of the first author on name only.

Sentiment analysis

Scores of sentiment and politeness of language use of each peer review report was performed using OpenAI's *ChatGPT* (GPT-3.5, Version Feb 13, 2023). The prompt consisted of the following question (see **Fig. 3a**):

Below you will find a scientific peer review. Such reviews generally contain the reviewer's sentiment in the first paragraph(s) of the review, followed by a list of specific recommendations to the authors. Can you score this peer review on [1] the sentiment, on a scale from -100 (negative) to 0 (neutral) to 100 (positive), and [2] politeness of language use, on a scale of -100 (rude) to 0 (neutral) to 100 (polite)? followed by the full text of the peer review. This question was entered into *ChatGPT* twice and the average of these scores was used for further analyses; for a correlation between the two iterations see **Supplementary Fig. 2**. Note that *ChatGPT* has become more reliable in recent updates, such that different iterations of scoring now produces a highly reproducible score (see **Supplementary Fig. 2**).

Statistics

To test the consistency across different reviewers of the same paper (**Aim 2; Fig. 3**), I used a combination of a mixed model, linear regression models and intra-class correlation coefficients. For **Fig. 3a** (differences between reviewers 1, 2 and 3), a mixed effects model was used to compute statistical significance, because repeated measures data was not always available (i.e., not all papers received a third review). This analysis was performed in Prism 9 (GraphPad Inc.). Linear regression and intra-class correlational analyses in **Fig. 3b** (sentiment scores across reviewers) and **Fig. 3c** (review scores vs. paper acceptance time) were performed using JASP 0.16 (University of Amsterdam). For the intra-class correlational analyses of **Fig. 3b**, ICC type ICC1,1 was used; because ICC is particularly sensitive to the assumption of normality, sentiment scores were first log transformed. For the polynomial linear regression in **Fig. 3c**, data were centered by *z*-scoring the individual sentiment and politeness scores.

To test the effects of author identity on review scores (**Aim 3; Fig. 4**), I used a combination of the Kruskal-Wallis ANOVA and Mann-Whitney tests. Note that review scores were not always normally distributed, so non-parametric tests were mostly used. To compute statistical significance in **Fig. 4a** (scores per field) and **Fig. 4b** (scores per institution location), a Kruskal-Wallis ANOVA was used. For **Fig. 4c** (scores correlated with 2023 QS World University Ranking), significance was calculated using linear regression. For **Fig. 4d and 4e**, Mann-Whitney tests were used to compute significance between male and female authors. Significant effects were further studied by splitting the reviews per score (i.e., splitting in lowest, median and highest scores per paper). To calculate statistical significance between male and female authors for lowest/median/highest score in **Fig. 4d and 4e**, Mann-Whitney tests were used. Statistical tests were always two-tailed. All analyses in **Fig. 4** were performed in Prism 9 (GraphPad Inc.). Significance was defined as $P < 0.05$ and denoted with asterisks; * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.

Data availability

All data are available as a Supplementary file to this paper.

Acknowledgements

J.P.H.V. was supported by a Rubicon postdoctoral fellowship from the Netherlands Organization of Scientific Research. I thank Amanda Tose, Han de Jong and Stephan Lammel for helpful comments on this manuscript.

Conflict of interest statement

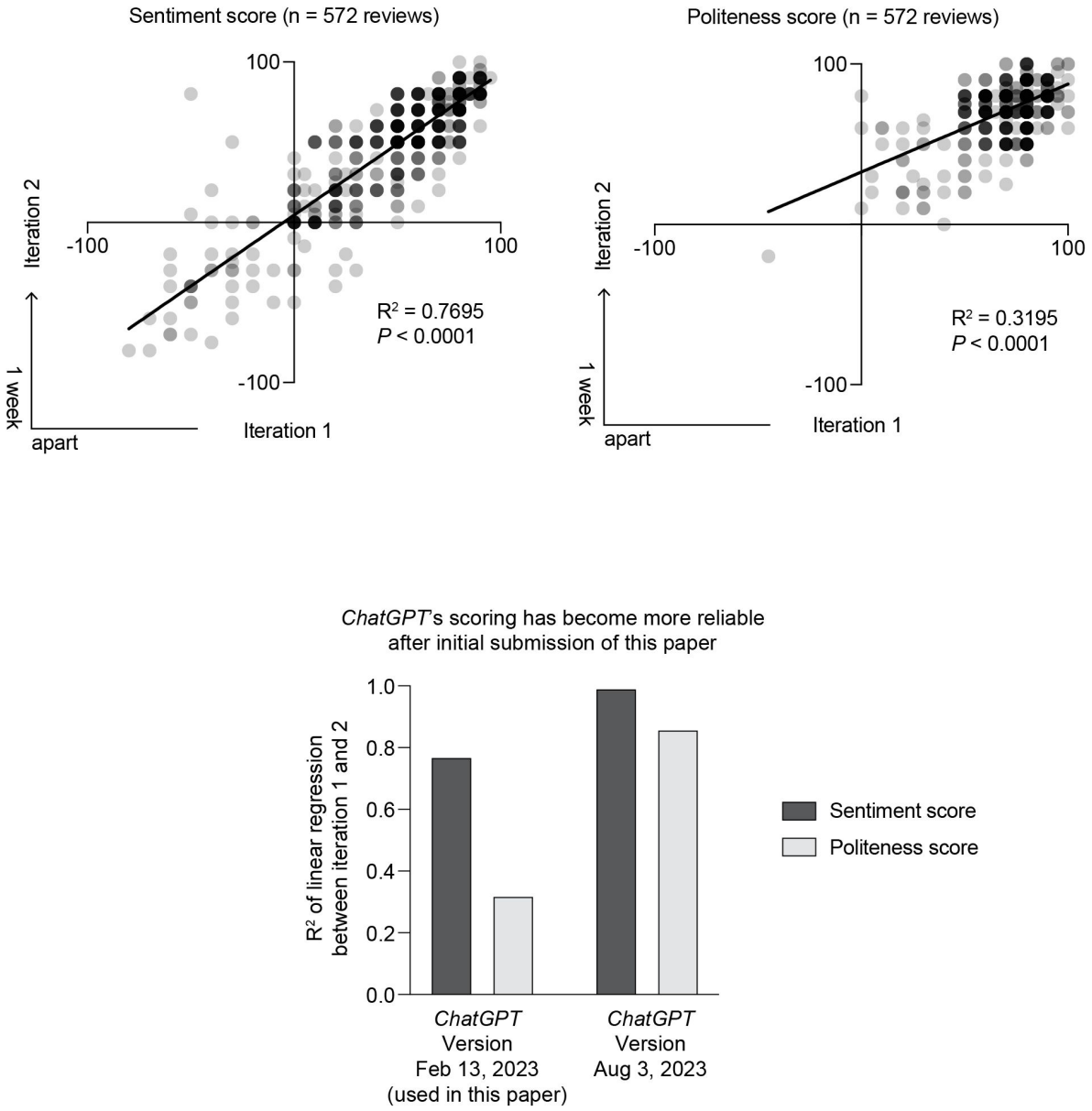
The author declares no competing interests.

Supplementary figure legends

Excerpt from review	*ChatGPT* analysis on entire review
From Shahin et al (Review #1) Overall there are some interesting findings in this study associated with the impact of loss of Ncf1 in NSCs. However, some of the presented data is phenomenological with conclusions that are unclear/equivocal. The claim that inactivating DNA repair promotes neurosphere forming ability (i.e. similar to WT cells) is surprising and it's difficult to understand why inhibiting key genome stability factors and increasing DNA damage (shown by H2AX formation) would promote recovery and be beneficial in terms of NSC function. While this might be a possibly interesting finding, as it stands there is too much uncertainty about what it means and how it happens. The data presented doesn't address how regulation of repair pathways modulates NSC function, if its physiologically relevant and the mechanistic nature of what DNA damage does to alter cell fate. (…)	Sentiment: -30 (slightly negative) Politeness: 20 (somewhat polite) The sentiment of this peer review is negative as the reviewer expresses several concerns about the clarity, relevance, and plausibility of the presented data and conclusions. The reviewer questions the significance and mechanism of the findings and highlights the limitations and inconsistencies of the study. However, the language use is not overly aggressive or impolite, and the reviewer provides constructive criticism and suggests possible ways to improve the study. Therefore, the politeness score is relatively neutral.
From Kang et al (Review #2) This manuscript reported a sequential narrow-field (SNF) stimulation method by applying sequential electrical stimulation on pairs of electrodes implanted in the hippocampus to control the localization of electric field. (…) This research is generally well designed and the manuscript is well organized. However, the idea of using multi implanted electrodeto delivery small electric current to control the activation volume is lack of novelty, even this method is designed for seizure control. There are also several major questions need to be resolved: (…)	Sentiment score: 10 Politeness score: 70 The sentiment of this peer review is somewhat positive, with a score of 10. The reviewer acknowledges the novelty of the research and praises the manuscript's organization and design. However, the reviewer also points out several major questions that need to be addressed, which lowers the sentiment score. The politeness score of this peer review is 70, indicating a polite and respectful tone. The reviewer uses professional language and structures their criticism in a constructive way, offering specific recommendations for improvement. The reviewer also acknowledges the strengths of the research and the manuscript, which shows a respectful attitude towards the authors.
From de Thonel et al (Review #1) This manuscript represents a major advance in understanding how mutations in a global protein acetylase, underlying a rare neurodevelopmental disorder, impacts cellular stress protection and neurodevelopment. (…) The authors present a very strong case for specific interactions between HSF2 within the KIX domain found in the coiled-coil region and CBP/p300 and validate the interaction with biochemical evidence, molecular modeling and mutagenesis. Moreover, while previous reports demonstrated the regulation of HSF2 via protein degradation, this work deciphers key mechanisms that regulate this process via protein lysine acetylation. Finally, the current work underscores the contributions of HSF2, and its dysregulation in RSTS, to stress-regulated gene expression and its impact in neuronal cell function. Overall, the results reported here are novel and highly impactful to the fields of stress regulation, neurodevelopmental processes and rare diseases.	Sentiment score: 90 Politeness score: 80 The sentiment of this peer review is very positive, with a score of 90. The reviewer praises the manuscript's rigor, strong case, and novel and impactful results. The reviewer also acknowledges the significance of the research in several fields, which raises the sentiment score. The politeness score of this peer review is 80, indicating a highly polite and respectful tone. The reviewer uses professional language and expresses admiration for the authors' work throughout the review. The reviewer's criticism is constructive and focuses on the strengths of the research, which shows a high level of politeness and respect towards the authors.
From Singleton et al. (Review #2) In another reanalysis alongside the 15+ other reanalyses of this dataset (more publications than participants), Singleton and colleagues provide partial evidence for the REBUS model. Although this may be one of the better tests of REBUS, the salami slicing has gotten out of control. Not only is the small sample (N = 15) not independent of all the other tests of REBUS, the analyses themselves are not necessarily independent (…) Although this paper should not be published in Nature Communications (or anywhere until all other datasets are integrated), wherever it may end up, the following points must be considered: (…) Not only am I not a fan of such fluffy writing that has become rampant in psychedelic publications, to act like this is some special feature of EBH is to ignore literally all of cognitive neuroscience. Nobody in cognitive neuroscience assumes dualism (at least explicitly). (…)	Sentiment score: -80 Politeness score: -45 The sentiment in this peer review is negative, with a score of -80. The reviewer is critical of the paper, particularly in relation to the small sample size and the "salami slicing" of the data. The reviewer also notes that some analyses are missing and recommends running k means clustering separately on placebo and LSD data. The language used is not particularly polite, with a score of -45. The reviewer uses language that is critical and accusatory, and does not make an effort to soften the tone of the review.

Supplementary Figure 1

Validation #1 — Examples of *ChatGPT* inputs and outputs



Supplementary Figure 2

Validation #2 — Consistency in sentiment and politeness scores for two different times *ChatGPT* was asked to analyze review reports.

(Top panels) A significant correlation was found between both times *ChatGPT* was asked to score the sentiment and politeness of the peer reviews. The average score of both iterations was used for all subsequent analyses in this paper. (Bottom panel) The relatively low R^2 values between both iterations shown in the top panels have greatly improved in versions of *ChatGPT* since the initial submission of this paper; R^2 values have now increased to 0.992 (for sentiment) and 0.859 (politeness).

- ORIGINAL REVIEW -

This paper aims to show that Area MT in monkeys has a direct correlation with the perception of motion. Several authors have already demonstrated that MT appears to be involved in 'driving' the perception of motion. The specific issue investigated here is whether MT (and particularly one cell type in MT) is involved in the perception of motion segmentation judgements. The most important finding is that a subset of cells in Area MT, which have been labelled pattern cells based on a single test, is particularly correlated with segmentation. The work is certainly of interest to people in the field. However, as the authors mention themselves, one of the largest problems in my mind is the importance of the finding in a broader context. Pattern and Component cells are defined based a on a very restricted stimulus-type: namely plaid patterns. Pattern and component cells do not form two distinct groups but, rather are defined based on a statistical test that splits cells into three groups that do not naturally split into three separate clusters.
There is no border between these groups (i.e. statistics don't allow you to identify three groups with dips between them). In truth, there is a continuum and most cells are a bit of both extremes. Plaid patterns contain restricted amounts of information on motion direction, so the lack of clear boundaries is not surprising. Therefore, saying that pattern cells are more closely linked to the perception of transparent motion appears philosophically to be a rather small increment. For example, what would happen if more certainty were built into the stimulus, such as occurs immediately, perceptually, when terminators appear in the RF? Would cells defined as pattern-selective differ from the other MT cells in this case and would this influence the bias of the single-cell responses? Would the 'special' case for pattern cells persist? The work was conducted in behaving primates and this is a very challenging and time consuming technique. To the best of my knowledge the work was conducted at the highest level and the statistics are appropriate. I believe that the experiments could be repeated by following the methods. I am not in a position to analyse the data to see if I can reproduce the same result. Therefore, while the work is of the highest quality, the paper suffers in my view from not clearly showing how the MT population code as a whole deals with transparency and segmentation judgements. I feel that a model that incorporates all of the response types in MT would be essential in trying to answer this. The authors could develop a data-driven model that uses their hard-won monkey data. As it stands, I feel that the paper attempts to highlight a correlation between perception and just one type of MT cell (i.e. pattern cells).

Sentiment: 10 Politeness: 75

Change	Examples of changes (highlighted sentences)	New scores
More positive	The work will be of great interest to people in the field. However, as the authors mention themselves, one small problem in my mind may be the importance of the finding in a broader context. The authors could develop a data-driven model that uses their hard-won monkey data. As it stands, I feel that this paper highlights a major advance between perception and one type of MT cell (i.e. pattern cells).	Sentiment: 67.5 (+57.5) Politeness: 60
More negative	The work is unlikely to be interest to people in the field. Indeed, as the authors mention themselves, an insurmountable problem in my mind is the importance of the finding in a broader context. The authors could develop a data-driven model that uses their hard-won monkey data. As it stands, I feel that the paper attempts to highlight an unimportant correlation between perception and just one type of MT cell (i.e. pattern cells).	Sentiment: -40 (-50) Politeness: 50
More polite	The work is certainly of interest to people in the field. Respectfully, as the authors mention themselves, one may doubt the the importance of their findings in a broader context. The authors could develop a data-driven model that uses their hard-won monkey data. Despite the author's hard work, I feel that the paper attempts to highlight a correlation between perception and one type of MT cell (i.e. pattern cells).	Sentiment: 50 Politeness: 90 (+15)
Less polite	I gotta say, some people will probably like this paper. However, as the authors mention themselves, their findings are just not that important. The authors must develop a data-driven model that uses their hard-won monkey data. Frankly, as it stands, the paper merely attempts to highlight a correlation between perception and just one type of MT cell (i.e., pattern cells). Not exactly groundbreaking, is it?	Sentiment: -20 Politeness: -55 (-130)

Supplementary Figure 3

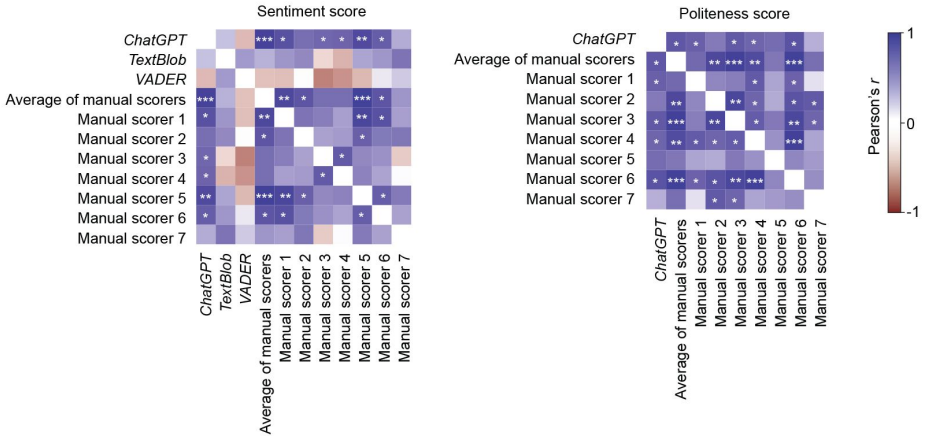
Validation #3 — Example of a manually manipulated review, showing that *ChatGPT* can pick up artificial changes in sentiment and language use.

It should be noted here that changes in sentiment usually also affected the politeness score (and vice versa), indicating that these scores are not fully independent. This may intuitively make sense; less polite language is often interpreted as more negative, also by human readers (see **Supplementary Fig. 4**).

	Sentiment									
	ChatGPT	Textblob[1]	VADER[2]	7 human scorers						
Paper				#1	#2	#3	#4	#5	#6	#7
Singleton et al	-80	7	99	Very negative	Very negative	Very negative	Negative	Very negative	Negative	Negative
Gaynes et al	-10	11	97	Negative	Negative	Very negative	Negative	Negative	Negative	Neutral
Quintana et al	30	8	-87	Positive	Very negative	Neutral	Neutral	Neutral	Neutral	Negative
Strembitska et al	40	14	94	Very positive	Positive	Negative	Negative	Positive	Positive	Positive
Audrain et al	72.5	10	100	Positive	Neutral	Positive	Positive	Neutral	Positive	Negative
Cicchini et al	75	10	100	Positive	Neutral	Negative	Positive	Positive	Positive	Positive
Bretheau et al	77.5	8	-97	Positive	Positive	Positive	Very positive	Positive	Neutral	Neutral

[1] 'Polarity' in Textblob ranges between [-1, 1]. This value was multiplied by 100 to match *ChatGPT*'s score range [-100, 100].
[2] 'Compound polarity score' in VADER ranges between [-1, 1]. This value was multiplied by 100 to match *ChatGPT*'s score range [-100, 100].

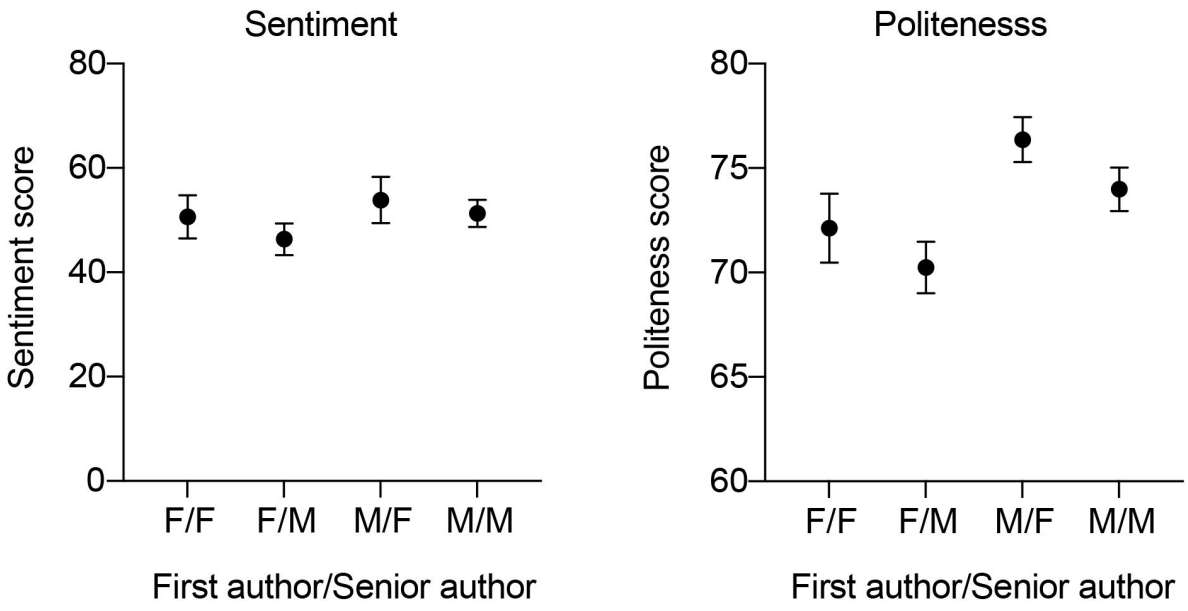
	Politeness									
	ChatGPT	Textblob	VADER	7 human scorers						
Paper				#1	#2	#3	#4	#5	#6	#7
Singleton et al	-32.5	n/a		Very rude	Very rude	Very rude	Very rude	Rude	Very rude	Neutral
Gaynes et al	50			Neutral	Very rude	Rude	Rude	Polite	Rude	Neutral
Bretheau et al	50			Neutral	Polite	Polite	Polite	Neutral	Polite	Very polite
Strembitska et al	62.5			Very polite	Polite	Polite	Polite	Neutral	Polite	Polite
Quintana et al	70			Polite	Rude	Neutral	Polite	Neutral	Polite	Rude
Audrain et al	77.5			Polite	Polite	Very polite	Very polite	Very polite	Very polite	Very polite
Cicchini et al	85			Neutral	Polite	Very polite	Neutral	Polite	Polite	Very polite



Supplementary Figure 4

Validation #4 — Comparison of *ChatGPT*'s scores of sentiment and politeness as compared to seven (blinded) human scorers for a diverse sample of reviews.

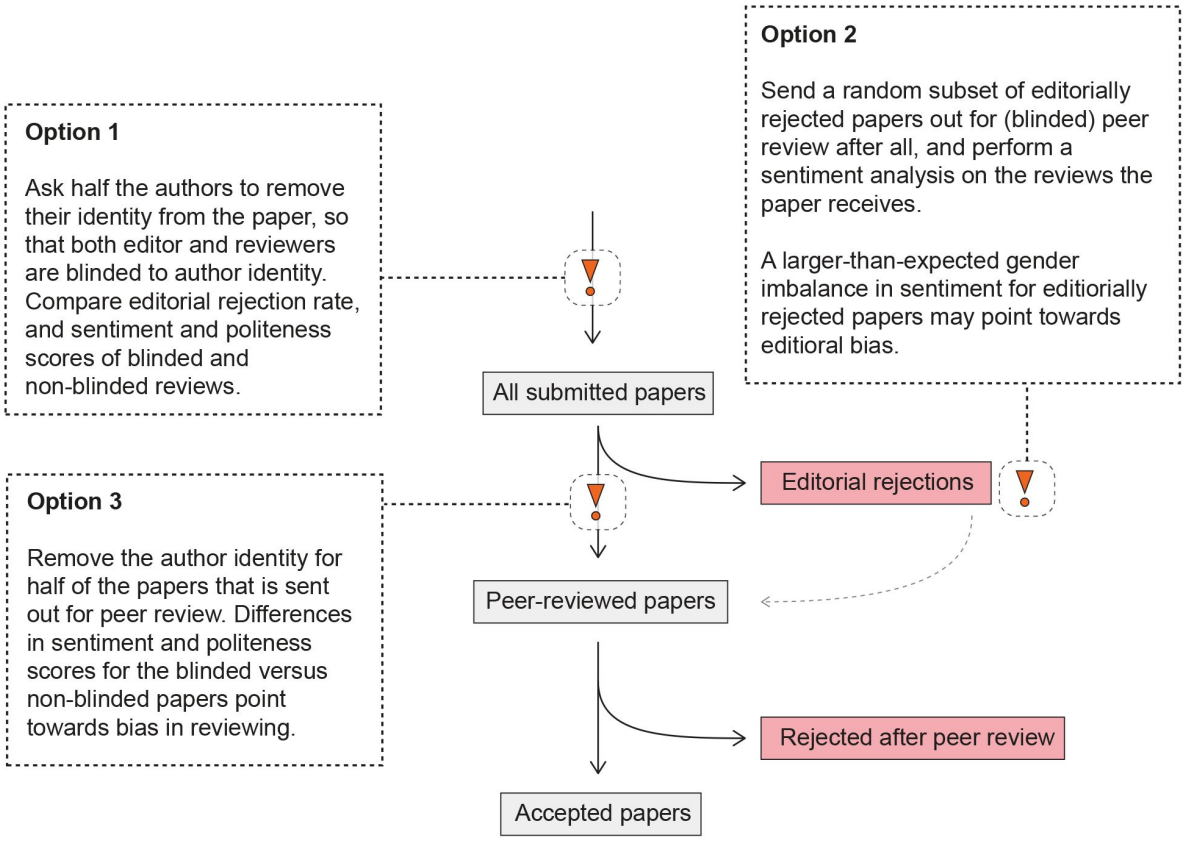
Figure showing high levels of variability across human scorers, but their average score had a high correlation with *ChatGPT*'s score. For this experiment, human scorers were asked to score the sentiment of seven reviews on a scale from very negative - negative - neutral - positive - very positive. Politeness was scored on a scale from very rude - rude - neutral - polite - very polite. For the cross-correlograms in the bottom panels, human scores were first converted to numbers on a scale from 1-5, so that these could be correlated to *ChatGPT*'s numerical scores. Scorers 1, 2 and 5 were scientists with a PhD; scorer 3, 4 and 6 were neuroscience graduate students; scorer 7 was a non-scientist. Please note that *TextBlob* and *VADER* do not provide politeness scores, and were thus not included in the second table. Asterisks indicate significance in a linear regression; * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$.



Supplementary Figure 5

Sentiment and politeness scores for papers in different gender groups.

As an example, 'F/M' indicates that the first author had a female name and that the senior author was male. Two-way ANOVAs on these data indicated no significant first × last author gender interaction effects for both sentiment ($F(1,397) = 0.05291$, $P = 0.8182$) and politeness ($F(1,397) = 0.02808$, $P = 0.8670$). $n = 73$ reviews for F/F, 126 for F/M, 44 for M/F, 158 for M/M.



Supplementary Figure 6

Experiments that journals can perform to rule out gender bias in reviewing and editorial decision making.

If these experiments show no evidence of bias, gender differences likely emerge higher up in the academic pipeline.

References

1. Park I., Peacey M.W., Munafò M.R. (2014) **Modelling the effects of subjective and objective decision making in scientific peer review** *Nature* **506**:93–6

2. Lipworth W.L., Kerridge I.H., Carter S.M., Little M. (2011) **Journal peer review in context: A qualitative study of the social and subjective dimensions of manuscript review in biomedical publishing** *Social Science & Medicine* **72**:1056–1063

3. King E.B., Avery D.R., Hebl M.R., Cortina J.M. (2018) **Systematic Subjectivity: How Subtle Biases Infect the Scholarship Review Process** *Journal of Management* **44**:843–853

4. 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**:2–17

5. Abramowitz S.I., Gomes B., Abramowitz C.V. (1975) **Publish or Politic: Referee Bias in Manuscript Review** *Journal of Applied Social Psychology* **5**:187–200

6. Blank R.M. (1991) **The effects of double-blind versus single-blind reviewing — experimental evidence from the American-Economic review** *The American economic review* **81**:1041–1067

7. Lundine J., Bourgeault I.L., Glonti K., Hutchinson E., Balabanova D. (2019) **"I don't see gender": Conceptualizing a gendered system of academic publishing** *Social Science and Medicine* **235**

8. Fox C.W., Meyer J., Aimé E. (2023) **Double-blind peer review affects reviewer ratings and editor decisions at an ecology journal** *Functional Ecology* **37**:1144–1157

9. Tomkins A., Zhang M., Heavlin W.D. (2017) **Reviewer bias in single-versus double-blind peer review** *PNAS* **114**:12708–12713

10. Chowdhary K.R. (2020) **Natural Language Processing. In: Fundamentals of Artficial Intelligence**

11. Hirschberg J., Manning C.D. (2015) **Advances in natural language processing** *Science* **349**:261–266

12. Yadav A., Vishwakarma D.K. (2020) **Sentiment analysis using deep learning architectures: a review** *Artificial Intelligence Review* **53**:4335–4385

13. Ghosal T., Kumar S., Bharti P.K., Ekbal A. (2022) **Peer review analyze: A novel benchmark resource for computational analysis of peer reviews** *PLoS ONE* **17**

14. Chakraborty S., Goyal P., Mukherjee A. (2020) **Aspect-based Sentiment Analysis of Scientific Reviews** *JCDL '20: Proceedings of the ACM/ IEEE Joint Conference on Digital Libraries in* **2020**:207–216

15. Luo J., Feliciani T., Reinhart M., Hartstein J., Das V., Alabi O., Shankar K. (2021) **Analyzing sentiment in peer review reports: Evidence from two science funding agencies** *Quantitative Science Studies* **2**:1271–1295

16. Transparent peer review for all (2022) **Transparent peer review for all. Nature Communications 13, 6173 (2022).** *Nature Communications* **13**

17. Loria S (2023) **TextBlob Documentation. Release v0.16.0**

18. Hutto C.J., Gilbert E.E. (2014) **VADER: A Parsiminous Rule-based Model for Sentiment Analysis of Social Media Text. Eight International Conference on Weblogs and Social Media (ICWSM-14)** *Ann Arbor, MI*

19. Liljequist D., Elfving B., Skavberg Roaldsen K. (2019) **Intraclass correlation - A discussion and demonstration of basic features** *PLoS ONE* **14**

20. Tennant J.P., Ross-Hellauer T. (2020) **The limitations to our understanding of peer review** *Research Integrity and Peer Review* **5**

21. Alam M., et al., vs Blinded (2011) **unblinded peer review of manuscripts submitted to a dermatology journal: a randomized multi-rater study** *Clinical and Laboratory Investigations* **165**:563–567

22. Snodgrass R. (2006) **Single-versus double-blind reviewing: an analysis of literature** *ACM SIGMOD Record* **35**:8–21

23. Blickenstaff J.C. (2006) **Women and science careers: leaky pipeline or gender filter?** *Gender and Education* **17**:369–386

24. Pell A.N. (1996) **Fixing the leaky pipeline: women scientists in academia** *Journal of Animal Science* **74**:2843–2848

25. Sheltzer J.M., Smith J.C. (2014) **Elite male faculty in the life sciences employ fewer women** *PNAS* **111**:10107–10112

26. Hengel E. (2022) **Publishing while female: Are women held to higher standards?** *Evidence from peer review. The Economic Journal* **132**:2951–2991

27. White K. (2010) **Women and leadership in higher education in Australia** *Tertiary Education and Management* **9**:45–60

28. Matías-Guiu J., García-Ramos R. (2011) **Editorial bias in scientific publications** *Neurología* **26**:1–5

29. Holmberg M.J., Andersen L.W. (2022) **Collider Bias** *JAMA* **327**:1282–1283

30. Griffith G.J., et al. (2020) **Collider bias undermines our understanding of COVID-19 disease risk and severity** *Nature Communications* **11**

Article and author information

Jeroen P. H. Verharen
Department of Molecular and Cell Biology and Helen Wills Neuroscience Institute, University of California, Berkeley, USA
For correspondence: jeroenverharen@gmail.com
ORCID iD: 0000-0001-7582-802X

Copyright

Editors

Reviewing Editor
Peter Rodgers
eLife, United Kingdom

Senior Editor
Peter Rodgers
eLife, United Kingdom

Reviewer #1 (Public Review):

Summary:
The author uses CHAT GPT in the assessment linguistic characteristics of peer reviews published from August 2022 to February 2023 in Nature Communications in neuroscience field. The author analysed over 500 reviews, which greatly varied in terms of author characteristics, peer review length, subfield, number of reviews and writing style. Chat GPT analysed reviews and gave the scores regarding the language characteristics related to sentiment score and politeness.

Strengths:
The innovative method is the biggest strength of this article. Moreover, the method can be implemented across fields and disciplines. I myself would like to see this method implemented in a grander scale. The author invested a lot of effort in data collection and I especially commend the that the chat GPT assessed the reviews twice, to ensure greater objectivity.

Weaknesses:
The weaknesses listed in my Public Review of the previous version have been addressed in this revised version.

Reviewer #2 (Public Review):

Summary
In this study, single author Jeroen Verharen investigates 500 publicly available peer review documents from 200 neuroscience papers. He uses ChatGPT to examine the sentiment and politeness of each review and performs a series of analyses including scores across reviewers, by field, institution ranking, and author gender. This is an impressive amount of analysis for a single author and uncovers an interesting pattern where female first authors receive consistently less polite reviews compared with male first authors. It is well known that women scientists face systematic discrimination across the field, and consistently in peer review. Using ChatGPT to examine these with a predefined scoring and metric system is novel and an accessible way for others in the future to evaluate these.
Strengths include:

1. Given the variability in responses from ChatGPT, he pooled two scores for each review and demonstrated significant correlation between these two iterations. He confirmed also reasonable scoring by manipulating reviews. Finally, he compared a small subset (7 papers) to human scorers and again demonstrated correlation with sentiment and politeness.
2. The figures are consistently well presented and informative. Figure 2C nicely plots the scores with example reviews. The supplementary data are also thoughtful and include combination of first/last author genders. It is interesting that first author female last author male has the lowest score.
3. A series of detailed analysis including breaking down reviews by subfield (interesting to see the wide range of reviewer sentiment/politeness scores in Computational papers), institution, and author's name and inferred gender using Genderize. The author suggests that peer review to blind the reviewers to authors' gender may be helpful to mitigating the impoliteness seen.
4. The author has strengthened the analysis in this revision by comparing it to lexicon- and rule-based algorithms TextBlob and VADER.

Weaknesses:
The weaknesses listed in my Public Review of the previous version have been adequately addressed in this revised version, and the article now acknowledges its limitations (ie, it is a pilot, proof-of-concept study, limited to articles about neuroscience). The author proposes further studies and it will be interesting to see the results of these.

Author Response

The following is the authors' response to the original reviews.

Summary of changes

I thank the reviewers for their thorough feedback on this paper and providing me with such a detailed list of recommendations. I have been able to incorporate many of their suggestions, which I believe has greatly improved this paper.

The most important changes:

• I added comparisons to the lexicon- and rule-based sentiment algorithms TextBlob and VADER to Supplementary Fig. 4. This shows the superiority of ChatGPT in scoring the sentiment of scientific texts compared to existing and already-validated tools for sentiment analysis based on natural language processing. [Suggestion Reviewer 2]

• I added the measure intra-class correlation to Fig. 3b, emphasizing the inconsistency in sentiment scores across different reviews of the same paper. [Suggestion Reviewer 3]

• I added Supplementary Fig. 6, in which I directly propose different experiments to test the causes of the observed gender effects on peer review. [Suggestion Reviewer 3]

• I further studied the issue of variability in responses by ChatGPT (Supplementary Fig. 2), and learned that this has greatly improved in the latest version of ChatGPT (for Version Aug 3, 2023, R2 values of 0.99 (sentiment) and 0.86 (politeness) were reached). I show these findings in Supplementary Fig. 2. [Suggestions Reviewers 1 and 3]

• Throughout the manuscript (most notably in the Abstract and Discussion), I emphasize that this is a proof-of-concept study, and make suggestions on how to scale this up across journals and fields. I also toned down certain claims given the relatively small sample size of this

study, including in the abstract. I also more prominently and elaborately discuss the limitations of the study in the Discussion section. [Suggestions Reviewers 1, 2 and 3]

• I made many smaller changes to text, figures and references on the basis of the reviewers' comments. [Suggestions Reviewers 1, 2 and 3]

Notably, Reviewer 3 has provided me with a very detailed list of recommendations for follow-up experiments. I appreciate their ideas, and I am currently considering different options for future work. Specifically I am looking to team up with a journal to perform the experiments laid out in Supplementary Fig. 6 of the new paper, to study whether I can find evidence of bias across rejected and accepted papers. As suggested by this reviewer, I am also looking into ways to automate data collection using APIs, and by utilizing the rapidly expanding databases for transparent peer review.

Based on this preprint, I have received messages from academics that are interested in using generative AI to study scientific texts. By revising this manuscript, I hope to provide them with the tools to concurrently expand the analysis of peer review into different scientific disciplines and journals.

> ### *Reviewer #1 (Public review)*
>
> *Strengths:*
>
> *The innovative method is the biggest strength of this article. Moreover, the method can be implemented across fields and disciplines. I myself would like to see this method implemented in a grander scale. The author invested a lot of effort in data collection and I especially commend that ChatGPT assessed the reviews twice, to ensure greater objectivity.*

I want to thank this reviewer for commending the innovative methodology of this study. I appreciate that this reviewer would like to see this methodology implemented at a grander scale, which is a view that I share. I initially only included Neuroscience papers, because I was uncertain whether I would be able to properly assess the reviews from different scientific disciplines (and thus judge whether ChatGPT was able to provide plausible scores).

The reviewers have provided me with a list of potential follow-up experiments, and I am currently considering different options for future work. Specifically I am looking to team up with a journal to perform the experiments laid out in (the new) Supplementary Fig. 6 of the new paper, to study whether I can find evidence of bias across rejected and accepted manuscript of a journal. In addition, as suggested by Reviewer #3, I am looking into ways to automate data collection using APIs, and by utilizing the rapidly expanding databases for transparent peer review. Importantly, based on this preprint, I have received messages from academics that are interested in using generative AI to study scientific texts. By revising this manuscript now, I hope to provide them with the tools to concurrently expand the analysis of peer review into different scientific disciplines and journals.

The comments I received from the different reviewers made me realize that I did not describe the intent of this paper well enough in the original submission. I rewrote much of the Abstract, to emphasize the proof-of-concept nature of this study, and rewrote the Discussion to focus more on the limitations of the study.

> *Weaknesses:*
>
> *I have several concerns regarding the methodology of the article. The first relates to the fact that the sample is not random. The selection of journal and inclusion and exclusion criteria do not contribute well to the strength of the evidence.*

Indeed, the inclusion of only accepted manuscript from a single journal is the biggest caveat of this paper. I have re-written much of the Abstract to emphasize that this is a proof-of-concept paper, hoping that other researchers concurrently expand this method to larger and more diverse datasets.

> *An important methodological fact is that the correlation between the two assessments of peer reviews was actually lower than we would expect (around 0.72 and 0.3 for the different linguistic characteristics). If the ChatGPT gave such different scores based on two assessments, should it not be sound to do even more assessments and then take the average?*

This was a great recommendation by this reviewer, and a point also raised by Reviewer #3. Based on their suggestion, I looked into how each additional iteration of scoring would reduce the variability of scoring for a subset of papers (thus being able to advice users on an optimal number of iterations).

Interestingly, I observed that ChatGPT has become significantly more reliable in providing sentiment and politeness scores in recent versions. For the latest version (ChatGPT Aug 3, 2023), R2 = 0.992 for sentiment and R2 = 0.859 for politeness were reached for two subsequent iterations of scoring. Unfortunately, OpenAI does not allow access to previous version of ChatGPT, so the current dataset could not be re-scored. Yet, based on these data, there may no longer be a need for people to perform repeated scoring. I show these data in Supplementary Fig. 2, as I believe this is very useful information for people who are interested in using this tool.

> ### *Reviewer #1 (Recommendations to author)*
>
> *I had some difficulties reading the article, so it would maybe help to structure the article more (e.g. In the introduction there are three aims stated, so the Statistical Analysis section could be divided in three sections, and instead of the link to figures, the author could state which variables were analysed in a specific manner) to be easier to comprehend the details. Also, I found on one place that the sample consisted of 572 reviews, and on other that it was 558.*

These are very good points. I re-wrote the statistical analysis for clarity (Page 7 of the manuscript). The 558 reviews was a mistake from my part, as I forgot to include the fourth review for the 14 papers that received four reviews in the histograms of Fig. 2b and the accompanying text. This has been updated.

> *For figures 1a and 1b it could be considered to enter the table instead of several figures.*

I thank the reviewer for pointing this out. I tried this suggestion, but I found it to reduce the readability of the paper. As an alternative, I now provide an Excel spreadsheet with all the raw data, so people can find all the characteristics of the included papers.

> *99.8% of the reviews analysed were assessed as polite. This is, in my opinion, extremely important finding, which shows that reviewers are still holding to certain degree of standards in communication, and it can be mentioned in the abstract.*

I very much agree with this reviewer; this has now been added to the Abstract.

> *In results you state that QS World Ranking is "imperfect" measure. When stating that in the results section, it poses the question why it is used in the study, so maybe it is more suitable for the discussion.*

This point is well taken. Even though the QS World Ranking score is imperfect, I still think it can be useful, as a rough proxy of perceived prestige of an institution. I now removed this "imperfect measure" statement from the Results section, and moved it to the Discussion (Page 5).

> *In the Results section, instead of using only p values, please add measures of effect (correlations, mean differences), to make it easier to place in the context.*

For the significant effects of Fig. 4, I have added these to the figure legends. Please note that the used statistical tests are non-parametric, so I reported the Hodges-Lehmann differences (which is the median of all possible pairwise differences between observations from the two groups).

> *I think the results interpretation should be softened a bit, or the limitations of the study should be placed as the second paragraph in the discussion, since this was only specific journal with specific subfield.*

I agree with this reviewer that the relatively small sample size of this paper demands more careful wording. Throughout the manuscript, I have toned down claims, and emphasized the "proof of concept" nature of this study (for example in the Abstract). I also moved the limitations section to the second paragraph of the Discussion, and elaborate more on the study's caveats.

> *Methods:*
>
> *The measure Review time was assessed from submission to acceptance, but this does not need to be review time since it takes a lot of time sometimes to find reviewers. that needs to be stated as the limitation.*

This point is well taken. I changed this to "Paper acceptance time" in Fig. 3 and the accompanying text.

> *Gender name determination methods differed between the assessment of the first authors and the last authors, and that needs stronger explanation.*

I appreciate this reviewer raising this point, which has also been raised by Reviewer #3. For this paper, I have carefully weighed the pros and cons of automated versus manual gender determination. Initially, my intention was to rely only on a programmatic method to identify authors' names. However, I came to realize that there were inaccuracies in senior author gender predictions made by ChatGPT/Genderize. This was evident to me due to my personal familiarity with some of these authors, either because they are famous or through personal interactions. It seemed problematic to me to proceed with this analysis knowing that these misclassifications would introduce unnecessary variability to the dataset.

The advantage of the relatively small sample size in this study was the opportunity to manually perform this task, rather than being fully dependent on algorithms. While I attempted manual gender identification for the first author as well, this was way more challenging due to their limited online presence. The discrepancy in gender identification accuracy between first and senior authors did not go unnoticed, and I acknowledge the issue it presents. I also recognize that, unlike senior authors, reviewers may not necessarily be familiar with the first authors of the papers they evaluate, as indicated in the original submission of this paper. In light of this, I sought input from several PIs who often serve as reviewers. Their feedback confirmed that they typically possess knowledge of senior authors' identities, for example through conferences, whereas the same is not true for first authors.

Yet, this may be different for other scientific disciplines, where the pool of reviewers might be bigger.

Notably, for future studies I may make a different decision, especially when I use larger datasets that require me to automate the process.

I also realize that my rationale for the different methods of gender determination was not explained well enough in the original submission; I now explain my reasoning more elaborately on Page 7 on the manuscript.

> *For sentiment analysis: Please state based on what the GPT made a decision? Which program? (e.g. for gender it used genderize.io)*

This has been added to Page 7.

> *Finally, your entire analysis can be made reproducible (since everything is publicly available). You can share ChatGPT chats as online materials with variables entered with the dataset analysed and the code. This would increase the credibility of the findings.*

I will make the entire raw dataset available through the eLife website, including all reviews and their scores.

> ### *Reviewer #2 (Public review)*
>
> *Strengths include:*
>
> 1. *Given the variability in responses from ChatGPT, the author pooled two scores for each review and demonstrated significant correlation between these two iterations. He confirmed also reasonable scoring by manipulating reviews. Finally, he compared a small subset (7 papers) to human scorers and again demonstrated correlation with sentiment and politeness.*
>
> 1. *The figures are consistently well presented and informative. Figure 2C nicely plots the scores with example reviews. The supplementary data are also thoughtful and include combination of first/last author genders. It is interesting that first author female last author male has the lowest score.*
>
> 1. *A series of detailed analysis including breaking down reviews by subfield (interesting to see the wide range of reviewer sentiment/politeness scores in computational papers), institution, and author's name and inferred gender using Genderize. The author suggests that peer review to blind the reviewers to authors' gender may be helpful to mitigating the impoliteness seen.*

Thank you.

> *Weaknesses include:*

> 1. *This study does not utilize any of the wide range of Natural Language Processing (NLP) sentiment analysis tools. While the author did have a small subset reviewed by human scorers, the paper would be strengthened by examining all the reviews systematically using some of the freely available tools (for example, many resources are available through Hugging Face [https:// huggingface.co/blog/sentiment-analysis-python (http://huggingface.co/blog/sentiment-analysis-python)]). These methods have been used in previous examinations of review text analysis (Luo et al. 2022. Quantitative Science Studies 2:1271-1295). Why use ChatGPT rather than these older validated methods? How does ChatGPT compare to these established methods? See also: colab.research.google.com/drive/ (http://colab.research.google.com/drive/) 1ZzEe1lqsZIwhiSv1IkMZdOtjPTSTlKwB?usp=sharing*

This was a great recommendation by this reviewer, and I have tested ChatGPT against TextBlob and VADER, the two algorithms also used by the Luo et al. study — see Supplementary Fig. 4. Perhaps unsurprisingly, these algorithms performed very poorly at scoring sentiment of the reviews. Please note that I also tested these two algorithms at scoring individual sentences, Tweets and Amazon reviews, which it did very well (i.e., the software package was working correctly). Thus, ChatGPT is better at scoring scientific texts than TextBlob and VADER, likely because these algorithms struggle with finding where in the review the sentiment is conveyed. I now discuss this on Pages 1, 3 and 4 of the manuscript.

> 1. *The author's claim in the last paragraph that his study is proof of concept for NLP to analyze peer review fails to take into account the array of literature already done in this domain. The statement in the introduction that past reports (only three citations) have been limited to small dataset sizes is untrue (Ghosal et al. 2022. PLoS One 17:e0259238 contains over 1000 peer review documents, including sentiment analysis) and reflects a lack of review on the topic before examining this question.*

I thank this reviewer for pointing me to this very useful study. I regret missing this one in my initial submission; I now discuss this paper in Pages 1 and 5 of the manuscript.

> 1. *The author acknowledges the limitation that only papers under neuroscience were evaluated. Why not scale this method up to other fields within Nature Communications? Cross-field analysis of the features of interest would examine if these biases are present in other domains.*

I share this reviewer's opinion that it would be very interesting to expand this analysis to different subfields. I initially only included Neuroscience papers, because I was uncertain whether I would be able to properly assess the reviews from different scientific disciplines (and thus judge whether ChatGPT was able to provide plausible scores). The different reviewers have provide me with a list of potential follow-up experiments, and I am currently considering different options for future work, including expanding into different fields within Nature Communications. Additionally, I am looking to team up with a journal to perform the experiments laid out in (the new) Supplementary Fig. 6 of the new paper, to study whether I can find evidence of bias across rejected and accepted manuscript papers of a journal. I am also looking into ways to automate data collection using APIs, and by utilizing the rapidly expanding databases for transparent peer review. Yet, based on this preprint, I have received messages from academics that are interested in using generative AI to study scientific texts. By revising this manuscript now, I hope to provide them with the tools to concurrently expand the analysis of peer review into different scientific disciplines and journals.

The comments I received from the different reviewers made me realize that I did not describe the intent of this paper well enough in the original submission. I rewrote much of the Abstract, to emphasize the proof-of-concept nature of this study, and rewrote the Discussion to focus more on the limitations of the study.

> ### *Reviewer #3 (Public review)*
>
> *Strengths:*
>
> *On the positive side, I thought the use of ChatGPT to score the sentiment of text was novel and interesting, and I was largely convinced by the parts of the methods which illustrate that the AI provides broadly similar sentiment and politeness scores to humans who were asked to rank a sub-set of the reviews. The paper is mostly clear and well-written, and tackles a question of importance and broad interest (i.e. the potential for bias in the peer review process, and the objectivity of peer review).*

Thank you.

> *Weaknesses:*
>
> *The sample size and scope of the paper are a bit limited, and I have written a long list of recommendations/critiques covering diverse aspects including statistical/inferential issues, missing references, and suggestions for other material that could be included that would greatly increase the usefulness of the paper. A major limitation is that the paper focuses on published papers, and thus is a biased sample of all the reviews that were written, which prevents the paper properly answering the questions that it sets out to answer (e.g. is peer review repeatable, fair and objective).*

I very much appreciate this reviewer taking the time to provide me with such a detailed list of recommendations. Below, I will respond to this list in a point-by-point manner.

> ### *Reviewer #3 (Recommendations to author)*
>
> *My main issues with the paper are that it is not very ambitious, and gave me the impression the aim was to write the first paper using ChatGPT to address this question, rather than to conduct the most thorough and informative investigation that would have been feasible (many obvious questions that could be addressed are not tackled, since the sample size is small and restricted). There are also issues with selection bias, and the statistical analysis, that have possibly led to erroneous inferences and greatly limit what conclusions can be drawn from the analysis. I hope my comments of use in further improving the paper.*
>
> *The repeatability of ChatGPT when calculating the two linguistic characteristics is low. Taking the average of multiple assessments is one way to deal with this. To verify that taking the average of, say, 5 scores gives a repeatable score, the author could consider calculating 10 scores for a set of 20-30 reviews, calculating two scores for each review using the first 5 and second 5 ChatGPT ratings, and then calculating repeatability across the 20-30 reviews. It is important to demonstrate that ChatGPT is sufficiently repeatable for this new method to be useful.*
> *Also, it might be possible to automate this process a bit to save time - e.g. the author could change the ChatGPT prompt, like "please rate the politeness of this review from -100 to +100, do it 10 times independently, and print your 10 ratings as well as their average". Hopefully the AI is smart enough to provide 10 independently-computed*

> *ratings this way, saving the need to copypaste the prompt into the chat box 10 times per review.*

This was a great recommendation by this reviewer, and a point also raised by Reviewer #1. Based on their suggestion, I looked into how each additional iteration of scoring would reduce the variability of scoring for a subset of papers (thus being able to advice users on an optimal number of iterations). I also tested this Reviewer's suggestion to ask ChatGPT to score many times, and give separate scores for each iteration — this worked very well.

Interestingly, I observed that ChatGPT has become significantly more reliable in providing sentiment and politeness scores in recent versions. For the latest version (ChatGPT Aug 3, 2023), R2 = 0.992 for sentiment and R2 = 0.859 for politeness were reached for two subsequent iterations of scoring. Unfortunately, OpenAI does not allow access to previous version of ChatGPT, so the current dataset could not be re-scored. Yet, based on these data, there may no longer be a need for people to perform repeated scoring. I show these data in Supplementary Fig. 2, as I believe this is very useful information for people who are interested in using this tool.

> *To my mind, the main reason to use an AI instead of one or more human readers to rank the sentiment/politeness of peer reviews is to save time, and thereby allow this study to have a larger sample size than would be feasible using human readers. With this in mind, why did you choose to download only 200 papers, all from the discipline of Neuroscience, and only from Nature Communications? It seems like it would be relatively easy to download papers from many more journals, fields of research, or time periods if using AI-based methods, and in fact it would have been feasible (though fairly laborious) for one person to read and classify the sentiment of the reviews for 200 papers.*
>
> *As well as providing more precise estimates of the parameters you are interested in (e.g. the consistency of reviews, and the size of the difference in reviewer sentiment between author genders), expanding the sample beyond this small set of papers would allow you to address other interesting questions. For example, you could ask whether the patterns observed for neuroscience are similar to those in other research disciplines, whether Nature Comms is representative of all journals (given there are other journals with public reviews), and you could test whether the male-female differences have become greater or smaller over time (e.g. by comparing the male-female differences observed in the past to the effect size observed in 2022-23). Additionally, the main analyses in this paper would have higher statistical power - for example, you only include 53 papers with a female senior author, giving you quite low power/ precision to estimate the gender difference in the average sentiment of reviews (given the high variance in sentiment between papers).*

I want to thank this reviewer for taking the time about possible ways to increase the impact of this work. I agree, these are all great suggestions, and there are many possibilities to apply ChatGPTbased natural language processing to scientific peer review. Respectfully, I chose to continue with publishing this work in the form of a proof-of-concept paper, because I currently do not have the resources to perform this (quite labor intensive) study. Below I will explain my reasoning, that I also shared with Reviewers #1 and #2.

I initially only included Neuroscience papers, because I was uncertain whether I would be able to properly assess the reviews from different scientific disciplines (and thus judge whether ChatGPT was able to provide plausible scores). The different reviewers have provide me with a list of potential follow-up experiments, and I am currently considering different options for future work, including expanding into different fields within Nature Communications. Additionally, I am looking to team up with a journal to perform the experiments laid out in (the new) Supplementary Fig. 6 of the new paper, to study whether I can find evidence of bias across rejected and accepted manuscript papers of a journal. I am

also looking into ways to automate data collection using APIs, and by utilizing the rapidly expanding databases for transparent peer review. Yet, based on this preprint, I have received messages from academics that are interested in using generative AI to study scientific texts. By revising this manuscript now, I hope to provide them with the tools to concurrently expand the analysis of peer review into different scientific disciplines and journals. The comments I received from the different reviewers made me realize that I did not describe the intent of this paper well enough in the original submission. I rewrote much of the Abstract, to emphasize the proof-of-concept nature of this study, and rewrote the Discussion to focus more on the limitations of the study.

> *Also, if you could include some reviews of papers that were reviewed double-blind, you could test whether the gender-related differences in peer reviews are ameliorated by double-blind reviewing. Nature Comms (and many other journals with open review) do have some double-blinded papers, and there is evidence that that double-blinding is preferentially selected by authors who think they will experience discrimination in the peer review process (DOI: 10.1186/s41073-018-0049-z), and also that double-blinding does ameliorate bias (DOI: 10.1111/1365-2435.14259), so this seems very relevant to the ideas under study here.*
>
> *I note that the PLOS journals allow open peer review, and there is an API for PLOS which one can use to download the reviews for a given paper (e.g. try this query to get to the XML file of a paper which has open peer review: http://journals.plos.org/plosone/article/file?id=10.1371/journal.pone.0239518&type=manuscript). Using an API could allow this project to be scaled up, because you can programmatically search for the papers with open reviews, download those reviews using the API and some code, and then score them using the same ChatGPT-based methods used for Nature Comms. Also, Publons recently merged with Web of Science (Clarivate), and you can now read all the open peer reviews on Web of Science for papers which had open review (e.g. for this paper: https://www-webofscience-com.napier.idm.oclc.org/wos/woscc/fullrecord/WOS:000615934800001). It would be possible to write to Web of Science, request access to their data or search engine, and programmatically download many thousands of papers and their associated reviews, and then use ChatGPT or a similar AI to score them all (especially if you can pass the reviews to ChatGPT for scoring programmatically, instead of manually copy-pasting the reviews into the chat box one at a time as it appears was done in the present study).*

These are great suggestions, and I have different plans for follow-up studies, including the use of APIs to download large batches of peer reviews. The analyses in this paper have been performed in February of this year, even before the ChatGPT API had been released, which did not let me automate the process at that time. As a result, these analyses have been performed manually. I realize that the field is moving rapidly, and that there are now different options to scale this up quickly.

I plan on using the suggestions from this Reviewer for follow-up experiment in a next paper, and publish this revision as a proof-of-concept paper. In this way, different researchers can optimally use ChatGPT-based sentiment analyses for similar studies without a delay.

> *As you acknowledge, there is a selection bias in this study, since you only include papers that were ultimately published in Nature Comms (missing reviews of papers that were rejected). This is a really big limitation on the usefulness of some of your analyses. For example, you found no relationship between author institutional prestige and reviewer sentiment. This could be evidence of a fair and impartial review process (which seems unlikely!), or it could be a direct result of selection bias (specifically a "collider bias", like the famous example involving height and skill among professional basketball players). The likelihood that a paper is published is positively related both to its quality and the*

> *prestige held by the authors, we might expect a flatter (or even negative) correlation between prestige and reviewer sentiment among papers that were published than among the whole set of papers (like how the correlation between height and speed/skill is less positive among NBA players than among the general population, since both height and speed/skill provide advantages in basketball).*

I agree with this reviewer that the selection bias is a major limitation of this study. I rewrote much of the Abstract and Discussion to tone down claims, and more prominently discuss the limitations of this study. I also made several suggestions for follow-up experiments.

> *In the section "Consistency across reviewers", you write that there was little similarity between review sentiment scores from different reviewers from the same paper, and then write "This surprising result indicates high levels of disagreement between the reviewers' favorability of a paper, suggesting that the peer review process is subjective." However I disagree with this conclusion for three reasons:*
>
> - *Firstly, your dataset only includes papers that were published, and thus there is a selection bias against manuscripts where both/all reviewers disliked the paper - the removal of this (probably large) set of reviews will add a (potentially very strong) downward bias to your estimate of how consistent the review process is (since you are missing all those papers where the reviewers agreed). I think that one cannot properly answer the question "are reviewers consistent in their appraisals" without having access to papers that were rejected as well as those that were accepted.*

I agree with this reviewer that there is a selection bias in this study, which I acknowledged throughout the initial submission of this manuscript. Indeed, having access to reviews of rejected papers will greatly increase my confidence in this finding. However, if there is consistency across reviewers in the entire pool of (post-review rejected+accepted) manuscripts, some of that has to trickle down into the pool of accepted papers. The correlation between sentiment scores of the different reviewers is so strikingly low (or even absent) that I simply cannot envision a way in which there is consistency across reviewers in the pre-editioral decision stage. Yet, I realize that this point is debatable. Therefore, I changed the phrasing of the Discussion section, including the following sentence:

That being said, the extremely low (or even absent) relation between how different reviewers scored the same paper was striking, at least to this author.

> - *Secondly, the method used to assess whether the reviews for each paper tend to be similar (shown in Figure 3b) does not fully utilize the information contained in the data and could be replaced with another method. (In the paper 3 univariate regressions compare the sentiment scores for R1 vs R2, R1 vs R3, and R2 vs R3, which needlessly splits up the data in the case of papers with more than 2 reviewers, reducing power.) You could instead calculate the intraclass correlation coefficient (aka 'repeatability'), to determine what proportion of the variance in sentiment scores is between vs within papers (I suggest using the excellent R package rptR for this). Note that the sentiment scores are not normally distributed, and so regular regression (as you used) or one-way ANOVA (which you might be tempted to use for the ICC calculation) are not ideal - consider using a GLM or transformation (the rptR package automates the tricky calculation of repeatability for generalized models).*

I thank this reviewer for pointing me towards this option. I added this analysis to Fig. 3b, which confirmed the inconsistency in sentiment scores for reviews of the same paper (ICC =

0.055). As suggested by this reviewer, I decided to perform the ICC on log-transformed data, as ICC calculation is very sensitive to non-normally distributed data.

> - *Thirdly, an alternative and very plausible hypothesis for this lack of similarity (besides peer review being highly subjective) is that ChatGPT is estimating the "true sentiment" of a review (i.e. what the reviewer intended to say) with some amount of error (e.g. due to limitations/biases in the AI, or reviewers struggling to make themselves understood due to issues such as writing in a second language, typos, or writing under time pressure), which dilutes the similarly in the estimated sentiment of the reviews. In other words, if the true sentiment values are strongly correlated, but there is random error in how those values are estimated by ChatGPT, then the correlation between reviewer scores for each paper will tend to zero as the error tends to infinity. Furthermore a nebulous quality like "sentiment" cannot be fully summarised in a single variable running from -100 to +100, and if you had used a more multi-dimensional classification system for the reviews (or qualitative assessment by human readers) you might have found that there is a bit more correspondence (I'm speculating here, but I think you cannot really exclude this and the paper doesn't mention this limitation).*

This point is well taken. I added caveats to the Discussion section on Page 5. Altogether, after taking these caveats into account, I do believe that this analysis convincingly demonstrates subjectivity in the peer review of this subset of papers. That said, I hope that my re-written discussion and additional analysis have added the necessary nuance to this point.

> *In Figure 3C, you write "Contribution of paper scores to review time". This strongly implies to the reader that the sentiment scores inferred for the reviews have a causal effect on the review time. This is imprecise writing (since the scores were calculated by you after the papers were published, and thus cannot be causal - you mean that the* actual reviews *affected the review time, not the scores), but more importantly you cannot infer any causality here since your dataset is observational/correlational. You could fix this by re-phrasing to emphasise this, e.g. "Statistical associations between paper scores and review time".*

This is a very good point raised by this reviewer. I have corrected the phrasing so it no longer implies causality.

> *For the analysis shown in Figure 4d and Figure 4e, I am not certain what you mean by "data split per lowest/median/highest sentiment score". This is ambiguous, and I am also not sure what the purpose of this analysis is or what it shows - I suggest re-writing for greater clarity (and ideally providing the code used in all your analyses) and perhaps revising the analysis. Additionally, an important missing piece of information from this analysis (and most analyses in the paper) is the effect size. For example, you don't report what is the difference in politeness score and sentiment score between male and female authors, and what is the SE and 95% CIs for this difference. From eyeballing the figure, it looks like the difference in politeness is about 4 points on your 200point scale - this is small in absolute terms, but might be quite large in relative terms given that "politeness score" usually hovered around a small part of the full 200-point scale. What is this as a standardised effect size (i.e. in terms of standard deviations, as captured by effect sizes like Cohen's d and Hedges' g)? Calculating this (and its 95% CIs) would allow you to say whether the difference between genders is a "big effect", and give an idea of your confidence in your effect size estimate and any inferences drawn from it. You even discuss the effect size in your discussion, so it would help to calculate the standardised effect size. If you're not familiar with effect size and why it's useful, I found this paper*

> *very instructive: https://onlinelibrary.wiley.com/ doi/abs/10.1111/j.1469-185X.2007.00027.x*

I agree with this reviewer that this phrasing was ambiguous. I now rephrased this on Page 4 of the manuscript:

To study whether these more impolite reviews for female first authors were due to an overall lower politeness score, or due to one or some of the reviewers being more impolite, I split the reviews for each paper by its lowest/median/highest politeness score. I observed that the lower politeness scores for first authors with a female name was driven by significantly lower low and median scores (Fig. 4d, bottom panel). Thus, the least polite reviews a paper received were even more impolite for papers with a female first author.

I also added effect sizes of the significant effects from Fig. 4 to its figure legend. Please note that the used statistical tests are non-parametric, so I reported the Hodges-Lehmann differences (which is the median of all possible pairwise differences between observations from the two groups).

> *"Double-blind peer review has been debated before, but has come under scrutiny for various reasons" - this is vague and unhelpful. I think it's worthwhile to properly engage with the debate and the substantial body of evidence in your paper, given your main focus is on potential bias in the review process based on authors' identities (e.g. gender, institutional prestige).*

I thank the reviewer for pointing this out. I rephrased this sentence to indicate that there is evidence that it helps to remove certain forms of bias (Page 5):

To address this issue, double-blind peer review, where the authors' names are anonymized, could be implemented. Evidence suggests that this is useful in removing certain forms of bias from reviewing8,9, but has thus far not been widely implemented, perhaps because some studies have cast doubt on its merits21,22.

I have also added a Supplementary Fig. 6 to this paper, in which I lay out how my tool can be used to study bias by applying it to single- and double-blinded reviews (see also my answer to the other question about this topic below).

> *On a related note, in the first paragraph, when discussing the potential of single-blind review to allow reviewers to essentially discriminate against papers by women, there is a key missing citation. This year, the first truly experimental test of this hypothesis was published (DOI: 10.1111/1365-2435.14259); a journal conducted a randomised controlled trial in which submitted manuscripts were reviewed either single- or double-blind. They found no effect of author gender on reviewer ratings or editorial decisions (though there was an effect of review type on success rate of authors from different countries). It would be better to cite this instead of reference 6, which as you acknowledge is methodologically flawed. This paper is also worth a read given your focus on Nature journals: DOI: 10.1186/s41073-018-0049-z.*

This point is well taken. I now cite this paper (citation #8) and rephrased this part of the Introduction (Page 1).

> *"Another - arguably more simple - solution [compared to double-blind peer review] could be for reviewers to be more mindful of their language use." Here, you seem to be saying that we don't need to blind author names during peer reviewers, because it would simpler if all reviewers were simply nicer! I object to this because A) double-blind review is easy to implement, and greatly reduces the opportunity to tune the review to the*

> *author's identity (and there is some experimental evidence that it works in this regard), and B) it seems like wishful thinking to say that we don't need to implement measures that reduce the scope for bias, because all reviewers could instead stop using impolite language.*

This is a very valuable comment. I rephrased this to emphasize that this is an additional measure.

> *"reviewers may want to use ChatGPT to extract a politeness score for their review before submitting" Yes, that's an interesting idea, and I can imagine that some (probably small) proportion of reviewers will be interested in doing this. But I think you should think bigger about wholesale changes to the review system that are possible because of AI like ChatGPT. For example, the submission platforms where reviewers submit their reviewers (e.g. ScholarOne, Manuscript Central) could be updated to use AI to pre-screen draft reviews, and issue a warning to reviewers, like "Our AI assistant has indicated that the writing in this review might be impolite (*example phrases here*) - would you like to edit your review before you submit it?" Also, reviewcredit platforms like Publons could display not only the number of reviews that someone wrote, but an AI-generated assessment of how constructive, detailed, and polite their reviews are (this would help nudge people into writing better reviews, and also give credit where it's due to careful reviewers, which is part of the aim of Publons and similar platforms). This is just off the top of my head - there are many other good ideas about how AI could transform the peer review process. Indeed, AI is already good enough to generate quite useful peer reviews and constructive criticism of draft papers, and will surely get better at this... this surely has lots of implications for science publishing over the coming decades.*

These are great suggestions for implementation of this tool. I now end the first paragraph of the Discussion (Page 4) with the following sentence:

Such an automated language analysis of peer reviews can be used in different ways, such as afterthe-fact analyses (as has been done here), providing writing support for reviewers (for example by implementation in the journal submission portal), or by helping editors pick the best papers or most constructive reviewers.

> *"Further research is required to investigate the reasons behind this effect and to identify in what level of the academic system these differences emerge." Here you could mention what this research would be - I think you'd need the full sample of reviewed papers, not just those that were accepted. Spell out what analyses would be required to test and falsify the various (very plausible and interesting) competing hypotheses that you mention for the male-female difference in sentiment scores.*

Great point. I added a Supplementary Fig. 6, in which I show a visual depiction of the experiments that can be performed to answer these questions.

> *"areas of concern were discovered within the academic publishing system that require immediate attention. One such area is the inconsistency between the reviews of the same paper, highlighting the need for greater standardization in the peer review process." I disagree here. I think it is natural for there to sometimes be differences in how two or more reviewers rate the quality of a paper, even if the peer review process were carefully standardised (e.g. via the use of a detailed "peer review form", which helps guide reviewers to comment on all important aspects of the paper - some journals use these). This is because reviewers differ in their experience, expertise, or interests, and so some reviewers will catch mistakes that others miss, or request stylistic changes that others*

> *would not. More broadly, it's often not possible to write a version of the paper that satisfies all possible reviewers.*

I re-phrased part of the Discussion on Page 5 to indicate other sources of inter-reviewer variability. Specifically, I mention that some variability in sentiment can be expected based on the different backgrounds of the reviewers:

Notably, some level of variability may be expected, for example due to different backgrounds, experiences, and biases of the reviewers. In addition, ChatGPT may not always reliably assess a reviews sentiment, adding some spurious inter-reviewer variability.

Yet, as also mentioned in my response to one of the previous questions, I still find the the extremely low levels of consistency striking, even after taking these possible sources of interreviewer variability into account.

> *"the maximum score an institution could receive was 100 (in 2023 this was Massachusetts Institute of Technology)" - this seems unnecessary information (just mention the score runs from 0-100).*

I agree with this reviewer that this was unnecessary information. This has been removed.

> *"reviewers are generally familiar with the senior author of papers they review and thus are likely aware of their gender identity." This seems like a strong assumption, and you don't provide any evidence for it Speaking personally, as a reviewer and journal editor I am often not familiar with the senior author, or I* am *familiar with the first author - I am not sure how often I know the senior author but not the first author or vice versa. It's also not always the case that the first author is a junior scientist and the last author a senior, famous one, as you imply. I suggest that you use the same approach to score the gender of both author positions, namely inferring their gender programmatically from their name (I agree that generally the important thing for the purposes of this study is the gender that reviewers will infer from the name, not the author's actual gender, and so gender estimation from first names is the correct approach).*

I appreciate this reviewer raising this point, and I have carefully weighed the pros and cons of both approaches. Initially, my intention was to rely only on a programmatic method to identify authors' names. However, I came to realize that there were inaccuracies in senior author gender predictions made by ChatGPT/Genderize. This was evident to me due to my personal familiarity with some of these authors, either because they are famous or through personal interactions. It seemed problematic to me to proceed with this analysis knowing that these misclassifications would introduce unnecessary variability to the dataset.

The advantage of the relatively small sample size in this study was the opportunity to manually perform this task, rather than being fully dependent on algorithms. While I attempted manual gender identification for the first author as well, this was way more challenging due to their limited online presence. The discrepancy in gender identification accuracy between first and senior authors did not go unnoticed, and I acknowledge the issue it presents. I also recognize that, unlike senior authors, reviewers may not necessarily be familiar with the first authors of the papers they evaluate, as indicated in the original submission of this paper. In light of this, I sought input from several PIs who often serve as reviewers. Their feedback confirmed that they typically possess knowledge of senior authors' identities, for example through conferences, whereas the same is not true for first authors. Yet, this may be different for other scientific disciplines, where the pool of reviewers might be bigger.

Notably, for future studies I may make a different decision, especially when I use larger datasets that require me to automate the process. I now more elaborately explain why I made this decision on Page 7 of the manuscript.

> *In the Abstract, you write "suggesting a gender disparity in academic publishing". This part of the sentence contains no information about what you think is the cause of the male/female difference, and no further interpretation of its ramifications, so I think you can just remove it (because "disparity" just means a difference, so you are effectively saying something redundant like "there was a difference between papers with male and female senior authors, suggesting there is a difference")*

I thank the reviewer for pointing this out. I replaced the latter part of this sentence with “(…) for which I discuss potential causes.”, which I think is better than a short summary of potential causes which may lack the nuance that such a topic deserves.