Can AI be used ethically to assist peer review?

As the rate and volume of academic publications has risen, so too has the pressure on journal editors to quickly find reviewers to assess the quality of academic work. In this context the potential of Artificial Intelligence (AI) to boost productivity and reduce workload has received significant attention. Drawing on evidence from an experiment utilising AI to learn and assess peer review outcomes, **Alessandro Checco, Lorenzo Bracciale, Pierpaolo Loreti, Stephen Pinfield, and Giuseppe Bianchi,** *discuss the prospects for AI for assisting peer review and the potential ethical dilemmas its application might produce.*

The scholarly communication process is under strain, particularly because of increasing demands on peer reviewers. Manuscript submissions to peer-review journals are growing roughly 6% annually. Every year, over 15 million hours are spent on reviewing manuscripts previously rejected and then resubmitted to other journals. Many of these could be avoided at the pre-peer review screening phase.

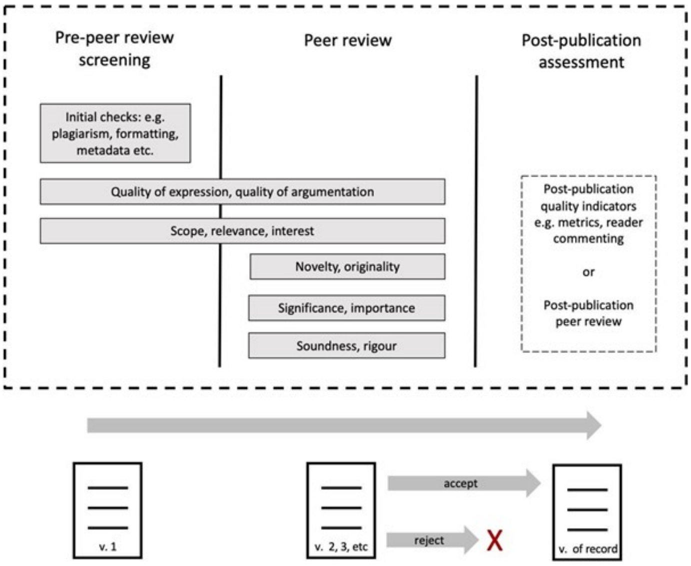
Pre-peer review screening
Peer review
Post-publication assessment
Initial checks: e.g. plagiarism, formatting, metadata etc.
Quality of expression, quality of argumentation
Scope, relevance, interest
Novelty, originality
Significance, importance
Soundness, rigour
Post-publication quality indicators e.g. metrics, reader commenting
or
Post-publication peer review
v. 1
v. 2, 3, etc
accept
reject
X
v. of record

Fig.1 Stages of the Peer Review process.

Rather than more grandiose visions of replacing human decision-making entirely, we are interested in understanding the extent to which AI might assist reviewers and authors in dealing with this burden. Giving rise to the question: can we use AI as a rudimentary tool to model human reviewer decision making?

Experimenting with AI peer review

To test this proposition, we trained a neural network using a collection of submitted manuscripts of engineering conference papers, together with their associated peer review decisions.

The AI tool analysed the manuscripts using a set of features: the textual content, together with readability scores and formatting measures. Our analysis covers the parts of the quality assurance process of outputs where pre-peer-review screening and peer review itself overlap, covering aspects like formatting, and quality of expression.

Once the learning phase was completed, we evaluated how accurate the empirical rules were in predicting the peer review outcome of a previously unobserved manuscript. Finally, we asked "*Why has the AI tool marked papers as accepted or rejected?"*, as answering that question may give us insight into the human decision-making the tool was modelling.

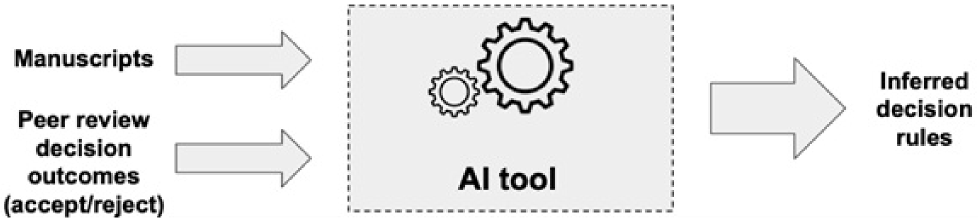

Fig.2 Schematised AI peer review tool.

Opening up AI decision making

Explaining models depending on half a million parameters is practically impossible using standard tools. Outcomes from the model can be affected by a whole range of different issues, such as the presence of specific words or particular sentence structures. We used a technique known as LIME to help explain what the model was doing in the case of specific documents. The technique is based on slightly changing the content of a document and observing how the model predictions change.

In the Fig.3 an example of an explanation for an accepted paper is shown. In orange, the top features influencing the decision towards a positive outcome are represented, while the blue colour represents factors associated with a negative decision. The absence of the word "quadratic", a low sentence count, and a high number of difficult/unusual words positively affects the model score, while a low number of pages, a small number of average syllables per word and a low text length affect the model score negatively. In some cases, explanations like this can expose potential biases or overfitting of the model: when the dataset is too small, the model could for example give too much importance to the presence/absence of a keyword.

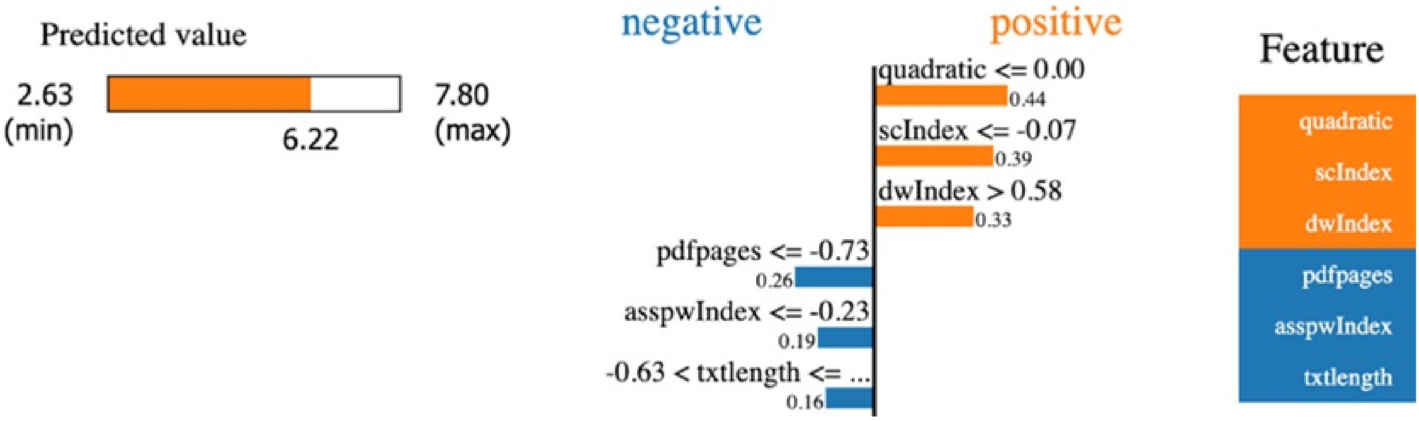

Fig.3 Explanation for machine learning based peer review decision.

Perhaps surprisingly, even using only rather superficial metrics to perform the training, the machine learning system was often able to successfully predict the peer review outcome reached as a result of human reviewers' recommendations. In other words, there was a strong correlation between word distribution, readability and formatting scores, and the outcome of the review process as a whole. Thus, if a manuscript was well written, used appropriate terminology and was well presented, it was more likely to be accepted.

One possible explanation for the success of this rather simplistic model is that if a paper is presented and reads badly, it is likely to be of lower quality in other, more substantial, ways, making these more superficial features proxy useful metrics for quality.

> Tools of the kind we developed have the potential to be of direct benefit in assisting editors of journals and conference proceedings in decision making.

However, it may be that papers that score less well on these superficial features create a "first impression bias" on the part of peer reviewers, who then are more inclined to reject papers based on this negative first impression derived from what are arguably relatively superficial problems.

Reviewers may be unduly influenced by formatting or grammatical issues (or the use of methods that have been associated with rejected papers in the past) and become unconsciously influenced by this in their judgements of more substantive issues in the submission.

In that case, an AI tool which screens papers prior to peer review could be used to advise authors to rework their paper before it is sent on for peer review. This might be of particular benefit to authors for whom English is not a first language, for example, and whose work, therefore, may be likely to be adversely affected by first impression bias.

Opportunities and Shortcomings

Tools of the kind we developed have the potential to be of direct benefit in assisting editors of journals and conference proceedings in decision making. They have the potential to save the time of reviewers, when used as decision support systems. They could also be useful to authors, as we have suggested. In particular, they might:

Reduce desk rejects

By catching the 'first impression', the approach we have explored in this paper has the potential to detect superficial problems early, like formatting issues and quality of the figures. Authors could be made aware of such problems immediately without any further review, or the AI tool could be used to pre-empt/inform desk rejects.

Improve human decision making with data

By analysing review decisions via a data-driven predictor/classifier, it is possible to investigate the extent to which the complex reviewing process can be modelled at scale. An analysis of the human decision process through data analysis and AI replication could potentially expose biases and similar issues in the decision-making process.

Biases and Ethical issues

A number of potential ethical issues arise from this work. Machine learning techniques are inherently conservative, as they are trained with data from the past. This could lead to bias and other unintended consequences, when used to inform decision-making in the future. For example, papers with characteristics associated with countries historically under-represented in the scientific literature might have a higher rejection rate using AI methods, since automated reviews will reflect the biases of the previous human reviewers, and, for example, may not take account of rising quality of submissions from such sources over time. Biases might also be introduced by the fact that historically, editors have disproportionately selected reviewers from high-income regions of the world, while low-income regions are under-represented amongst reviewers, the tool may then reflect the biases of previous reviewers.

An author will not trust an automated review if there is no transparency on the rationale for the decision taken. This means that any tools developed to assist decision making in scholarly communication need to make what is going on under the bonnet as clear as possible. This is particularly the case since models are the result of a particular design path that has been selected following the values and goals of the designer. These values and goals will inevitably be "frozen into the code".

> An author will not trust an automated review if there is no transparency on the rationale for the decision taken

It is also worth noting that tools designed to assist reviewers can influence them in particular ways. Even using such tools only to signal potentially problematic papers, could affect the agency of reviewers by raising doubts in their minds about a paper's quality. The way the model interprets the manuscript could propagate to the reviewer, potentially creating an unintended biased outcome.

All of these ethical concerns need to be considered carefully in the way AI tools are designed and deployed in practice, and in determining the role they play in decision-making. Continued research in these areas is crucial in helping to ensure that the role AI tools play processes like peer review is a positive one.

This post draws on the authors' paper AI-assisted peer review, published in Humanities and Social Sciences Communication and is a collaboration between the University of Sheffield and the University of Rome "Tor Vergata".

Note: This article gives the views of the author, and not the position of the Impact of Social Science blog, nor of the London School of Economics. Please review our Comments Policy if you have any concerns on posting a comment below.

Image Credit: In text images reproduced with permission of the authors, featured image LSE Impact Blog.