ARTICLE

Check for updates

AI-assisted peer review

Alessandro Checco [1✉], Lorenzo Bracciale[2✉], Pierpaolo Loreti[2], Stephen Pinfield[1✉] & Giuseppe Bianchi[2]

The scientific literature peer review workflow is under strain because of the constant growth of submission volume. One response to this is to make initial screening of submissions less time intensive. Reducing screening and review time would save millions of working hours and potentially boost academic productivity. Many platforms have already started to use automated screening tools, to prevent plagiarism and failure to respect format requirements. Some tools even attempt to flag the quality of a study or summarise its content, to reduce reviewers' load. The recent advances in artificial intelligence (AI) create the potential for (semi) automated peer review systems, where potentially low-quality or controversial studies could be flagged, and reviewer-document matching could be performed in an automated manner. However, there are ethical concerns, which arise from such approaches, particularly associated with bias and the extent to which AI systems may replicate bias. Our main goal in this study is to discuss the potential, pitfalls, and uncertainties of the use of AI to approximate or assist human decisions in the quality assurance and peer-review process associated with research outputs. We design an AI tool and train it with 3300 papers from three conferences, together with their reviews evaluations. We then test the ability of the AI in predicting the review score of a new, unobserved manuscript, only using its textual content. We show that such techniques can reveal correlations between the decision process and other quality proxy measures, uncovering potential biases of the review process. Finally, we discuss the opportunities, but also the potential unintended consequences of these techniques in terms of algorithmic bias and ethical concerns.

[1] Information School, The University of Sheffield, Sheffield, UK. [2] Department of Electronic Engineering, University of Rome Tor Vergata, Rome, Italy.
✉email: a.checco@sheffield.ac.uk; Lorenzo.Bracciale@uniroma2.it; s.pinfield@sheffield.ac.uk

Introduction

The scholarly communication process is under strain, particularly because of increasing demands on peer reviewers and their time. Manuscript submissions to peer-review journals have seen an unprecedented 6.1% annual growth since 2013 and a considerable increase in retraction rates (Publons, 2018). It is estimated over 15 million hours are spent every year on reviewing of manuscripts previously rejected and then resubmitted to other journals (AJE, 2018).

Developments that can make the quality control/assurance process associated with research outputs, particularly the peer review process, more efficient are likely to be welcomed by the research community. There are already a number of initiatives making use of automated screening tools in areas such as plagiarism prevention, requirements compliance checks, and reviewer-manuscript matching and scoring. Many of these tools make use of artificial intelligence (AI), machine learning and natural language processing of big datasets. Some notable examples are:

- Statcheck, software that assesses the consistency of authors' statistics reporting, focusing on p-values (Nuijten et al., 2017).
- Penelope.ai, a commercial tool able to examine whether the references and the structure of a manuscript meet a journal's requirements.
- UNSILO, software able to automatically pull out key concepts to summarise manuscript content.
- StatReviewer, which checks the soundness of statistics and methods in manuscripts (Shanahan, 2016).
- Automated tools used in the grant-review processes of the National Natural Science Foundation of China, to reduce bias and the load on the selection panels (Cyranoski, 2019).
- Online system to manage the grant application process, introduced in 2012 by the Canadian Institutes of Health Research (CIHR), removing the need for face-to-face meetings, to reduce reviewer fatigue and improve quality, fairness and transparency.
- Automated Essay Scoring (AES) application, used by EdX, MIT and Harvard's non-profit MOOC federation to assess written work in their MOOCs.

Such initiatives are not without controversy, however. Some doubts have been expressed about the reliability of the Statcheck tool (Schmidt, 2017). The CIHR application system received heavy criticism from some reviewers (Akst, 2016). Other MOOC producers have been skeptical of the AES scoring application (Balfour, 2013).

It is, therefore, helpful to investigate further the potential of big data and AI to support the quality control process in general, and peer review process in particular, and investigate specifically how the time of peer reviewers might be saved, especially in the more tedious parts of the review process, which require less intellectual input or domain expertise. That is what we aim to do in this study.

Peer review is also under strain in the sense that it is coming under increasing scrutiny from those who are concerned that it may often reinforce pre-existing biases in the academy. Biases associated with gender, language or institutional affiliation are examples of those, which may be evident in decisions made within the peer review system (Lee et al., 2013). Such biases may arguably come to the fore, particularly if unconscious, when reviewers are time pressured and do not adequately reflect on their own decision-making. Investigation of the system using AI tools may help, therefore, not merely to save reviewers' time, but also to uncover biases in decision-making. Uncovering such biases may help to develop approaches to reducing or eliminating their impact.

The quality control/assurance process for research publishing typically consists of a number of different components, as delineated by Spezi et al. (2018). Their work focuses on peer-reviewed journals but applies to other quality-controlled research outputs, e.g., conference proceedings. They divide the normal process that takes place prior to publication in two stages:

- **Pre-peer review screening:** consisting of a number of checks, including plagiarism detection, formatting checks, scope verification etc, plus checking of language and quality of expression. In many cases, if a paper does not meet these checks, it will be "desk rejected" before peer review. However, the extent of pre-peer review screening varies considerably across different publications.
- **Peer review:** normally consisting of assessment of four main criteria: novelty or originality, significance or importance, relevance or scope, rigour or soundness. In addition, peer reviewers are also asked to comment on the quality of the language and argumentation (overlapping with but also extending the language checks carried out in the pre-peer review screening).

Spezi et al. (2018) also discuss various post-publication quality identifiers, including citation and usage metrics, and reader commenting, as well formal post-publication peer review, as carried out by, e.g., *F1000 Research*. Post-publication commenting and community-based analysis can in extreme circumstances result in retraction of articles where it becomes evident a study was flawed.

In Fig. 1, we recast the model developed by Spezi et al. to visualise the different dimensions of the peer review process. We continue to use the framework of pre-publication screening, pre-publication peer review, and post-publication quality indicators, but have attempted to show more clearly where criteria used in pre-screening and peer review intersect, the point that is the focus of the research presented in this paper.

Our research covers some aspects of the pre-peer review screening, particularly formatting, language and expression. Pre-peer review screening includes a variety of checks shown in Fig. 1, including formatting checks. Assessment is also made at the pre-screening phase of quality of expression and scope issues. Consideration of these issues is also undertaken by peer reviewers, who assess quality of expression and argumentation and issues of relevance and interest to a particular subject community as part
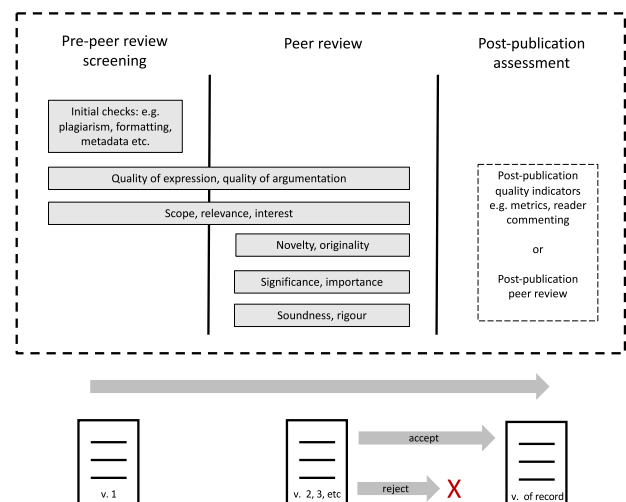

Fig. 1 Dimensions of the peer review process. Quality metrics and their relevance along the phases of the peer review process.

of their work. The submitted document may undergo several revisions during the process, and will then be formally accepted or rejected for publication. Published documents are normally fixed in the form of a "version of record".

These two phases are then followed by a post-publication phase (than can affect a paper's status, e.g., retraction).

Aim. Our goal is to make an early contribution to the discussion on the potential, pitfalls and uncertainties of the use of AI to assist pre-peer review screening as well as some of the aspects of the peer review process, based on the results of an empirical investigation aiming to reproduce outcomes of reviewer decision-making using AI methods.

We are interested in understanding the extent to which AI can assist reviewers and authors, rather than in attempting to replace human decision-making processes. At the same time, we are also interested in investigating the ways in which using AI as a rudimentary tool to model human reviewer decision-making can uncover apparent biases in the human decision-making process, and particularly, the extent to which human decision-making may make use of different quality proxy measures, which could produce inequitable assessments. Using AI tools to identify such biases could then help in addressing them.

More specifically, our research questions are:

RQ1: To what extent can AI approximate human decisions in the quality assessment and peer-review process?

RQ2: Can AI play a role in the decreasing time reviewers need to spend assessing papers?

RQ3: Can AI uncover common biases in the review process?

RQ4: What are the ethical implications of the use of such tools?

RQ1 is important since AI approaches may sometimes encounter major problems in trying to imitate abstract and complex intellectual activity, such as peer review, so their accuracy in modelling human decision-making needs to be carefully evaluated. RQ2 raises the possibility of AI tools being used to address some of the strains in the peer review process by potentially avoiding redundant reviews, and removing or at least reducing, the burden of standardised checking (AJE, 2018). RQ3 focuses on the extent to which (potential) biases may be evident in review outcomes, in particular how human decision-making may make use of proxy measures of quality, which may reflect bias, and how AI tools may uncover this. RQ4 is important since it encourages reflection on the ethical implications of using AI tools in assisting human decision-making, in particular whether such tools can help address issues of bias or, in fact, whether their use may even risk perpetuating bias.

We address RQ1 by performing and evaluating our experiment in sections "Methodology" and "Results" of this paper, while section "Explainability" reports the reasoning behind our model and its limitations. We address RQ2 in section "Impact.", where we show how AI can potentially reduce redundant reviews, administrative functions and standardised checks. RQ3 is addressed in sections "Analysis of the experiment outcome" and "Bias", and RQ4 in section "The ethics of (Semi) automated peer review", where we analyse the implications of the experiment. We observe that care needs to be taken in how AI tools are used in this space.

Our approach. In this paper, we focus on peer-reviewed conference proceedings, and report an experiment designed to investigate how well a neural network can approximate the known recommendations of peer reviewers. To do that, we trained the neural network using a collection of submitted manuscripts, together with their associated peer review decision (acceptance/rejection or average review score), as outlined in Fig.

Fig. 2 Schematic illustration of the training approach. Manuscripts and peer review decision outcomes are inputs of the AI tool to infer decision rules.

2. The AI tool analysed the manuscripts using a set of features: the textual content (word frequencies), together with readability scores (measuring factors such as word sizes, sentence lengths, and vocabulary complexity, indicating how easy it is to understand the text) and formatting measures—features that might be considered somewhat separate from the substance of the research reported in the document. The analysis therefore covers the parts of the quality assurance process where pre-peer-review screening and peer review itself overlap (Fig. 1), covering aspects of pre-peer-review checks, e.g., formatting, and peer review, e.g., quality of expression.

The objective underlying the use of the AI tool is to find a set of empirical rules to correlate a posteriori the document features with the final peer review decision. This approach is explained in more detail in section "Methodology".

Once the learning phase was completed, we evaluated how accurate such empirical rules were in predicting the peer review outcome of a previously unobserved manuscript. Finally, we examined the AI decision criteria to understand what a machine can learn about the quality assessment and peer review process (section "Explainability"). Asking "Why the AI tool has marked papers as accepted or rejected?" is particularly relevant where its "decision" correlates well with the recommendations of reviewers since it may give us insight into human decision-making.

Findings. 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. This correlation between simple proxy quality measures and the final accept/reject decision is interesting, and merits further discussion and investigation.

We discuss in section "Analysis of the experiment outcome" the significance of this finding, particularly in relation to what it tells us about the quality control process in general and the peer review process in particular.

Limitations of our approach. The approach we take does not cover all the aspects of the peer review process, nor does it attempt to replace human reviewers with AI. However, it suggests that there are *some components* of the quality assessment and peer review process, which could reasonably to assisted or replaced by AI-assisted tools. These could potentially include readability assessment of the text, and formatting checks, as well as more established checks, e.g., plagiarism detection. Conversely, we do not envisage any relevant contribution from AI on the processes requiring significant domain expertise and intellectual effort, at least for the foreseeable future. The possible model of quality assessment we are exploring is then a semi-automated one, where AI informs decision-making, rather than alone determining outcomes. It is acknowledged the extent to which this is possible will vary depending on a number of factors, not least the nature of the research output itself and its approach to presenting research

results. One key variable here is in disciplinary norms. There is likely to be considerable variation across different disciplines in the ways AI assessment tools can be designed and applied to research outputs.

Structure of this paper. The rest of this article is structured as follows. We first introduce related work, in particular studies on peer review and relevant aspects of AI. We then present our methodology, and discuss the accuracy and the explainability of the obtained models. Following that, we analyse the experiment results. We go on to discuss the significance of our findings, the applicability of the proposed system to publishing practice, and some of the key ethical implications. Finally, we draw conclusions and suggest possible future work.

Related work

The peer review process is complex, and itself takes place in a complex wider research system. Judgements of quality take place as part of a system "managed by a hyper-competitive publishing industry and integrated with academic career progression" (Tennant, 2018). It is a system that combines extensive collaboration with intense competition between academic researchers and institutions (Tennant, 2018). Nevertheless, the "invisible hand" of peer review is still considered to be what keeps the quality of refereed journal literature high (Harnad, 1999; Mulligan et al., 2013; Nicholas et al., 2015). While a future with different approaches to scholarly communication can easily be envisioned (Priem, 2013), it is hard to imagine one without peer review (Bornmann, 2011; Harnad, 1998).

Several studies have analysed how potentially "problematic publications" (e.g., those containing fraudulent research) may be identified through peer review and have provided good practice guidelines for editors (Horbach & Halffman, 2019). Problems with the peer review system have been observed focusing a wide variety of problems, ranging from the opportunistic (or even adversarial) rejection of high-quality work, to the acceptance of low-quality manuscripts without a careful review (D'Andrea & O'Dwyer, 2017).

A number of recent initiatives have experimented with major changes to the peer review process. Most notably, open peer review is being more widely introduced as an alternative paradigm of interaction between authors and reviewers (Ford, 2013; Ross-Hellauer, 2017). In the case of open-access mega-journals (OAMJs), the review policies are pared down to focus on rigour and soundness only, leaving to "the community to decide" on issues of novelty, significance and relevance following publication (Spezi et al., 2018). Other approaches have included quality judgements being made following publication, sometimes shifting ideas of peer review to potentially include post-publication commenting by readers (Pontille & Torny, 2015). A wide range of alternative peer review processes, systems and online solutions (from Reddit-like voting systems to block-chain models) are explored by Tennant et al. (2017).

While the number of studies of peer review systems is vast, less quantitative analysis of the actual process of reviewing manuscripts has been carried out. Piech et al. (2013) studied how to identify and correct for the bias of reviewers in Massive Open Online Courses (MOOCs). Some MOOCs have already started to use machine-based Automated Essay Scoring (AES) applications to assess work, although others have pointed out potential problems in using such tools (Balfour, 2013).

To understand how the peer review process may be supported by AI tools, an important precondition is understanding how the quality and readability in textual data can be assessed. Readability formulas and cognitive indices has been studied extensively

(Crossley et al., 2011, 2008), and Natural Language Processing (NLP) has proven to be a powerful tool for text quality assessment (Cozza et al., 2016). However, assessing the quality of complex documents by automated means is still a challenging problem (Sonntag, 2004).

Modelling of the peer review process has been attempted in other contexts, such as education research (Goldin & Ashley, 2011), and in legal education contexts (Ashley & Goldin, 2011), which may be relevant for our study.

One thing that is apparent, however, is that many socio-cultural biases are present in peer review (Lee et al., 2013), and some of them could potentially propagate to AI systems, as described in the studies on algorithmic bias (Garcia, 2016; Mittelstadt et al., 2016). Many studies have shown that biased algorithms can inadvertently discriminate against specific groups (Barocas & Selbst, 2016; Zarsky, 2016).

Bias in the review process may take different forms. These include "first-impression" bias, the Doctor Fox effect, ideological/theoretical orientation, language, perceived social identity and prestige biases (Hojat et al., 2003; Lee et al., 2013; Siler et al., 2015). Such biases are evident in many contexts, such as websites (Lindgaard et al., 2006; SWEOR, 2019), examinations (Wood et al., 2018) or staff recruitment (Florea et al., 2019). In the area of document assessment, the typographical layout has been proven to have an important role in the "first-impression bias", where initial impressions of the document colour further judgements about its overall quality (Moys, 2014). Challenges remain in modelling this, although there are some pioneering studies that show how AI techniques can be used to model first-impression bias in relation to human encounters, e.g. job interviews (Gucluturk et al., 2017).

As the peer review process is a highly complex and demanding set of tasks, we suggest that, especially when time is at a premium, reviewers may tend to employ heuristics (D'Andrea & O'Dwyer, 2017) to assess a paper (e.g., more superficial features of the document, like language, formatting, etc.). Such heuristics can potentially be approximated using AI. Indeed, recent trends demonstrate the ability of AI to approximate human cognition in some simple tasks in a way that is similar to the way humans use their senses to relate to the world around them (Russell & Norvig, 2016). However, understanding how far we are from machine approximation to the full task of peer review, with all of its complex intellectual input, is still an open question.

Methodology

To investigate how the review process works, it is necessary to have access to a set of submitted papers and their corresponding review reports (including the specific scores assigned by reviewers). This itself is not easy, as the reviews, and especially the content of rejected manuscripts are usually confidential. In section "Data collection". we explain in detail how we overcame this challenge in the data collection process we employed. Once a set of papers has been acquired together with their review scores, it is necessary to perform a set of transformations on the data to obtain relevant features. The process we carried to do this is described in section "Feature extraction". After that, some statistics on the documents also need to be collected to help the modelling process, as shown in section "Feature augmentation—macroscopic features". Finally, the features can be used to train a neural network, as described in section "Neural network design". We include a significant level of technical detail in this section for reasons of transparency and in order to enable the replicability of our experimental setup. The process we followed is represented in Fig. 2, a schematic of the training of the AI tool. By inputting both the submitted manuscripts and peer reviewer recommendations/

Table 1 Collected datasets summary. For WCNC only the average score is available.

	ICLR.cc/2018	ICLR.cc/2019	WCNC 2018
No. manuscripts	909	1414	1018
Average review score (training set)	5.45	5.43	3.01
Average review score (test set)	5.36	5.46	3.00
Minimum review score (training set)	2.0	1.5	na
Minimum review score (test set)	2.0	2.33	na
Maximum review score (training set)	9.0	8.67	na
Maximum review score (test set)	8.67	8.67	na
Accepted papers ratio (training set)	37.1%	35.6%	48.9%
Accepted papers ratio (test set)	36.3%	36.2%	47.8%
Number of words	89,372	134,724	110,930
Number of non-unique words	35,458	50,118	36,795

decisions into the AI tool, we were able to produce a set of inferred decision rules, which underpinned the decisions made.

Data collection. We employed two different strategies to obtain the review data. Firstly, we obtained submitted manuscripts, numerical reviewers' scores and editorial decisions from the general chair of the 2018 IEEE wireless communications and networking conference (WCNC). Secondly, we employed data from openreview.net (aka OpenReview), which provides "a flexible cloud-based web interface and underlying database API enabling [...] open peer review, access, discussion and publishing"[1]. We selected two conferences with the largest number of publications from openreview.net, that is the International Conference on Learning Representations (ICLR) for the years 2018 and 2019. All the papers from both ICLR and WCNC had been reviewed by two to five reviewers. We cleaned the data to remove any information that had been added after manuscript acceptance, e.g., author names and affiliations. For simplicity we did not use the textual reviews, but rather we focused only on the numerical scores of the review. In summary, for all of the datasets we had: a paper (pdf file), an editorial decision (accepted/rejected), and numerical reviewers' scores (e.g., 3.5).

In Table 1, a summary of the data is shown, together with the dimensions of the training set used to build the model and the test set used to evaluate its predictive capabilities. We can observe that the datasets are fairly balanced. Figs. 3 and 4 show that the separation between accepted and rejected papers in terms of score is quite apparent. However, in a small number of cases, there is a rather strong discrepancy between the editorial decision and the average score of the reviewers. This is particularly true for OpenReview data: after a preliminary analysis on the dataset, we did not find any rule or correlation that appreciably binds the final acceptance decision with the document content, apparently corroborating Colman (1982), who also observed a lack of correlation between paper quality, authors' reputation/affiliation and the final accept-reject decisions for journal papers. We did, however, identify some patterns when using the average reviewers scores (previous step of the review process). For this reason, we decided to focus on predicting the average reviewers score for OpenReview data.

Feature extraction. The pdf documents were converted to textual data. Then, each document text was tokenised[2] using binary encoding of the top 20,000 words in terms of frequency for the WCNC conference and term frequency-inverse document frequency (TFIDF) of the top 2000 words for OpenReview.

Feature augmentation—macroscopic features. As further discussed in sections "Related work" and "Analysis of the
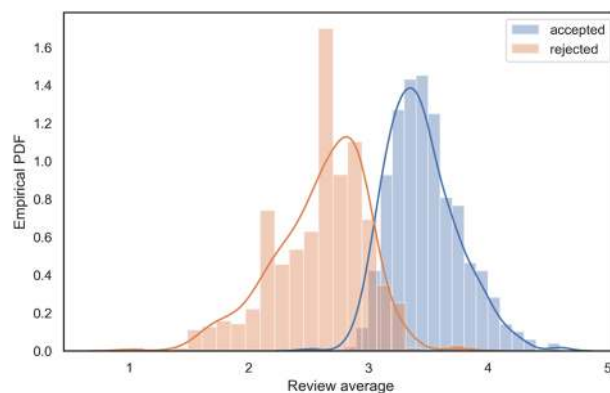

Fig. 3 WCNC distribution of average review score for accepted and rejected papers. Empirical probability distribution function of the average review score for accepted (in blue) and rejected (in orange) WCNC papers.
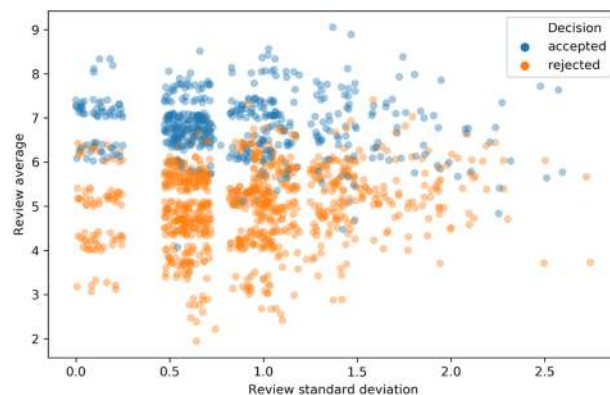

Fig. 4 Review standard deviation vs. review average for OpenReview datasets. Scatterplot of average vs. standard deviation of accepted and rejected OpenReview papers, with added jitter to improve visibility.

experiment outcome", the layout of the document and its graphical components could affect the first-impression of the reviewer, and thus are important in our modelling process. For this reason, we added "macroscopic features" like text/image ratio, file size and textual length to our analysis. We also incorporated the most commonly-used text quality and readability metrics (Crossley et al., 2011), as shown in Table 2.

Neural network design. For all the datasets, we used well-accepted standards of developing neural networks, technical details of which are given below. We employed a dense neural

network with a 32 neurons layer, followed by a dropout layer to reduce overfitting, feeding to a layer of 16 neurons. The two layers used Rectified Linear Unit (ReLU) activation. The final layer comprises a single node with a sigmoid activation function when the network is trained to predict paper acceptance or rejection or to predict the reviewer's scores. The resulting networks include 640,577 total trainable parameters for WCNC, 64,577 total trainable parameters for OpenReview. The difference is accountable to the different number of input features for the two analysed cases. The loss function was the binary cross entropy for the classification problem (WCNC) and the Mean Squared Error (MSE) for the regression problem (OpenReview). To train the network we made use of Stochastic Gradient Descent (SGD) with Nesterov update.

Aim. Using a standard machine-learning approach, we split the dataset in a training set on which the learning algorithm builds its model, and a test set used to evaluate the model accuracy. The model accuracy is defined as the ability to predict, respectively: (i) whether a previously unobserved paper would be accepted or not (for the WCNC dataset); (ii) the reviewers average score (for the OpenReview dataset).

Results

In this section, we show the results in terms of prediction performance of the designed models with respect to the final editorial decision. For an analysis of the models, see section "Analysis of the experiment outcome".

WCNC dataset. For this dataset, we used as baseline a random classifier, that would obtain an F1-score[3] of about 50% since the dataset is balanced. We measured accuracy, precision, recall and F1-score. Depending on the context in which the model is used, one of these metrics on its own might be more appropriate to assess the usability of the model. For example, recall might be the best measure if the tool was meant to signal problematic papers to assign additional reviewers: in that case, a false negative (a high-quality paper signalled as low quality) might create an additional burden on the reviewer and, at the same time, reduce the confidence on the tool. Conversely, F1-score might be more appropriate in assessing the quality of the prediction if the cost of false positives and false negatives are expected to be the same. The results are shown in Table 3.

By focusing on the first layer of the neural network, we can observe that the stronger features in activating the neurons are the Linsear write formula (Crossley et al., 2011), the text length and the number of pages, together with the following list of words: *address, approach, approximately, conclusion, constant, correlation, deployed, drawn, easy, efficient, illustrates, increased, issue, knowledge, level, obtain, page, potential, previously, process, respectively, types*, γ, τ. However, it is important to note that the interactions between the features are more complex than a simple independent activation, and involve multiple layers in the neural network. This is why we dedicate section "Analysis of the experiment outcome" to an extensive analysis of the interpretability of the model.

OpenReview dataset. As discussed before, here we focused on the prediction of the reviewer average score, using the Mean Absolute Error (MAE) and the Mean Squared Error (MSE) as reference metrics. As baseline we selected a naive classifier that chose the median score of the training dataset. In Fig. 5, the training behaviour of the network is shown, in terms of MSE of the validation and training set over the different training epochs. The performance of the trained regressor is shown in Table 4.

As we can see, the trained regressor is able to improve a naive one.

Even more importantly, Fig. 6 shows the empirical distribution of the MAE over the test set: we can see that 75% of the samples have an error of under 1.2 (over a total score of 10), and the median error is only 0.79. This means that we can expect this model to predict the average reviewers score with a median error of 0.79 over 10. While the reduction in the error rate is promising, it is worth noting that its low magnitude (for both approaches) is in part explained by the low variance of the scores used by reviewers, who tend not to use the whole scale available.

As for the previous dataset, we can observe that the stronger features activating the first neural layer are the Lix index, the Flesch Kincaid grade (Crossley et al., 2011), the text length, number of pages and file size, together with the following list of words: *256, actor, buffer, causal, coefficients, curve, demonstrate, dnn, exploration, github, gpu, idea, imagenet, measures, message, perturbations, precision, produce, quantised, query, regression, review, selected, sentence, standard, state, supervision, tensor, token, width*. We refer to the next section for a more detailed analysis of the explainability of the model.

Table 2 List of computed macroscopic features (Crossley et al., 2011).

Macroscopic feature	Shortcode
Automated Readability Index	ariIndex
Avg letter per word	alpwIndex
Avg character per word	acpwIndex
Avg sentence length	aslIndex
Avg syllables per word	asspwIndex
Char count	ccIndex
Coleman Liau index	cliIndex
Dale Chall readability score	dcrsIndex
Difficult words ratio	dwIndex
Flesch Kincaid grade	fkgIndex
Flesch reading ease	freIndex
Gunning fog	gfIndex
Letter count	lcIndex
Lexicon count	llcIndex
Linsear write formula	lwfIndex
Läsbarhetsindex (LIX)	lixIndex
Polysyllabcount	pscIndex
Anderson's Readability Index (RIX)	rixIndex
Sentence count	scIndex
Smog index	siIndex
Syllable count	sscIndex
Text length	txtlength
Number of pages	pdfpages
File size	pdfsize
Text/images ratio	textdensity

Table 3 WCNC classification performance vs. random classifier.

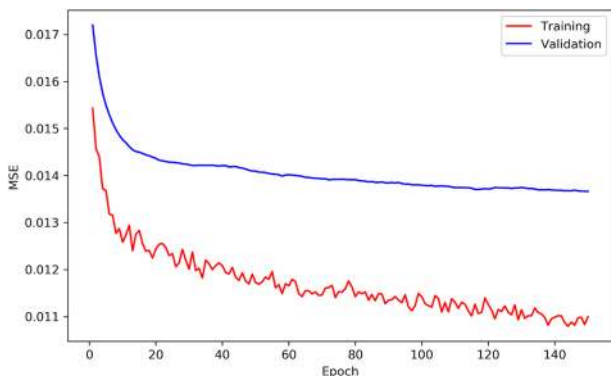| Classifier	Accuracy [%]	F1-score [%]	Precision [%]	Recall [%]
Random	~50%	~50%	~50%	~50%
Dense NN	74.01%	72.30%	72.45%	73.19%

Fig. 5 AI learning process on OpenReview dataset. Mean Squared Error vs. number of training epochs for OpenReview training process, with batches of 32 samples.

Table 4 OpenReview regression performance vs. naive regressor (the lower the better).

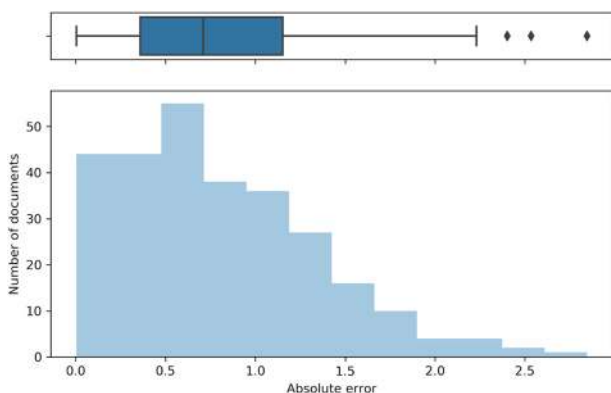| Regressor	MAE	MSE
Naive	0.96	1.40
Dense NN	0.79	0.90

Fig. 6 Absolute error empirical distribution. 75% of the samples have an error of under 1.2, and the median error is 0.79.

Explainability. In the case of black-box models like deep-learning systems, such as the one we developed, it is important to attempt to interpret the reasoning of the model, or in other words, the rationale for an automated decision, to allow practitioners to decide the level of trust given to a model. This is of fundamental importance to reduce the opacity of such tools, enabling an informed evaluation of their performance and, therefore, allowing greater trust in their outputs.

Explaining models depending on half a million parameters is practically impossible using standard tools. For example, the presence of a specific keyword or a specific document statistics can affect the model decision in a non-linear and document-dependent way, making it very challenging to identify a set of simple rules that can make the model understandable to a non-specialist.

However, recent studies have shown that local interpretable model-agnostic explanations (LIME) are able to effectively explain the model decision on a specific document (Ribeiro et al., 2016). The technique is based on the local perturbation of
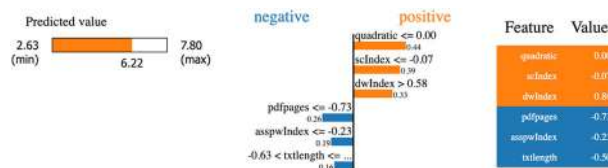

Fig. 7 LIME explanation for a document in the OpenReview dataset. The top features affecting the model are: the word "quadratic", the sentence count scIndex, the number of difficult words dwIndex, the number of pages pdfPages, the average number of syllables per word asspwIndex, and the text length txtlength.

an instance and the development of a linear model. We used this method to create explanations on multiple documents, and repeat this approach on the whole document space, with the goal of picking a set of exemplar documents distant enough from each other to obtain a representative set of rules of the whole model (submodular pick technique).

In Figs. 7 and 8 examples of the local explanation for an accepted paper of the OpenReview dataset are shown. In orange the top features influencing the decision towards a positive decision are represented, while the blue colour represents factors associated with a negative decision. This summary is simple enough to be presented as is to a non-specialist. In Fig. 7, the absence of the word "quadratic", a low sentence count, and a high number of difficult 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, the local explanation can expose potential biases or signal overfitting of the model. Overfitting occurs where the fits to the model is based closely on the specifics of the training set but would be a poor fit further related data. For example, in Fig. 8, we can observe that the absence of the word "decoding" is affecting negatively the model decision. This might reasonably be considered an overfitting problem caused by the over-abundance of documents related to decoding in this conference. The choice of whether this contingent explanation is satisfactory is highly context dependent, but it can increase the transparency of the model and allow the practitioners to assess the trust on the model. Another example of overfitting has been observed in the early stages of the model building, using a less than optimal hyperparameter set. In that case, the presence of some specific first names was regarded as a positive signal for the final decision.

Often local rules do not generalise for documents that are considerably different. For example, high text length can increase the predicted score when some keywords are present, while it could be decrease it in other contexts. This is clearly shown after running a submodular pick analysis of the whole space, as shown in Fig. 9, after identifying a group of exemplar documents covering the training space. Some keywords like "hyperparameters" and "quadratic" can be modelled as positive or negative depending on the context of the specific paper.

Analysis of the experiment outcome
Despite the focusing on rather superficial document features, like word distribution, readability and formatting scores, the machine-learning system was often able to successfully predict the peer review outcome. In other words, those documents that scored highly in those areas (e.g., they achieved high scores in readability and were formatted as required) were more likely to be recommended by reviewers for acceptance, and those that achieved lower scores in those areas, more likely to be recommended for rejection. There are a number of possible explanations for this.
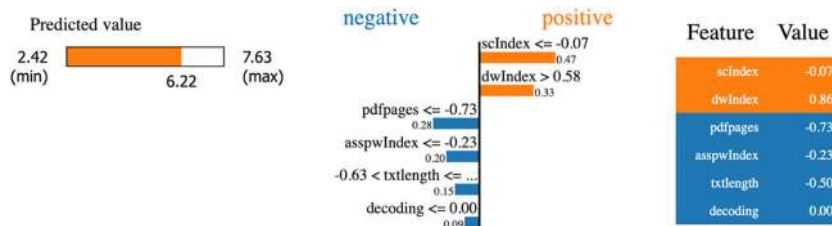
Fig. 8 LIME explanation for a different document in the OpenReview dataset. The top features are: the word "decoding", the sentence count `scIndex`, the number of difficult words `dwIndex`, the number of pages `pdfPages`, the average number of syllables per word `asspwIndex`, and the text length `txtlength`.
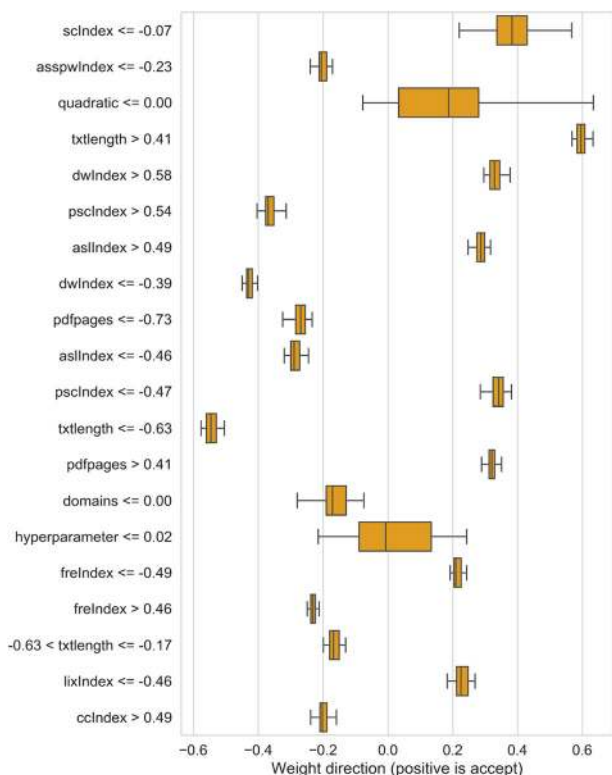

Fig. 9 Distribution of top 20 explanation features for rules covering the training space. Boxplot representing the weight direction for each feature (vertical black line is the median).

Table 5 Potential role of AI in the different dimensions of the peer review process.	
Dimension	**AI impact**
Formatting	High
Plagiarism	High
Scope	High
Readability/English	Medium
Relevance	Medium
Soundness/rigour	Low
Novelty	Low
Impact	Low

One possible explanation is that a correlation between such superficial features and the outcome of the review process as a whole might indicate that they are in fact a good indicator of the overall quality of the paper. In other words, 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. In that case, assessing a paper taking into particular account of those superficial features may be a reasonable heuristic in making overall decisions about the quality of the paper. If that was the case, it would be reasonable to use AI to screen papers before peer review process, using AI as a tool to desk reject papers based on these macroscopic features as part of the pre-peer review screening referred to earlier. This would save the time of peer reviewers who would have to review papers, which were highly likely to low quality. Even if low-scoring papers are not desk rejected, it could be that their scores are flagged to peer reviewers to assist them in making their decisions—also a potential time saver.

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, e.g., formatting or grammatical issues and become unconsciously influenced by this in their judgements of more substantive issues in the submission. Examples of such issues in papers include the presence of typos, the presence of references to papers from regions under-represented in the scientific literature, or the use of methods that have been associated with rejected papers in the past.

In that case, an AI tool that screens papers prior to peer review in the way described, could be used to advise authors to rework their paper before it is sent on for peer review, since it is likely that peer reviewers may reject the paper or at least be negatively influenced by the macroscopic features of the paper, which could be relatively easily corrected.

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.

Discussion

Impact. Table 5 lists several aspects and the potential role how AI-based tools, such as the one we describe in this study, can (or already do) impact the different dimensions of the peer review process.

Tools of this kind have the potential to be of direct benefit in assisting editors of journals and conference proceedings in decision-making (and similarly, the role of making funding decisions, as described in section "Conclusions and future work"). Such tools have the potential to save the time of reviewers, when used as decision support systems. We suggest there may be potential positive impacts in the following specific processes.

Reducing desk rejects. By catching the "first eye impression", the approach we have explored in this paper has the potential to detect early superficial problems, like formatting issues and quality of the graphs. Authors could be made aware of such problems immediately without any further review, or the outcome could be used to pre-empt/inform desk rejects. Even though this technique could wrongly signal high quality (but unusual) typographical choices, a notification about potential issues would help authors to evaluate whether or not they should correct their presentation. Removing superficial problems before peer review could help to avoid reviewer decisions being unduly informed by first-impression biases, and allow them to focus more on the scientific content. On the other hand, AI could also provide inexperienced reviewers with an impartial point of view of the work, providing some performance indicators and synthetic parameters such as a measure of deviation from past authors in terms of style, language and typographic format.

Explaining decisions by data-driven modelling. 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. Although complex (our preliminary neural network has half a million parameters), such models can be inspected to derive justifications for and explanations of decisions. While completely replicating the cognitive tasks required for the peer review process would be demanding, an analysis of the human decision process through data analysis and AI replication could potentially more easily imitate the more superficial parts of the decision-making processes involved. This could in turn potentially expose biases and similar issues in the decision-making process.

Discovering latent motivations. Motivations behind a decision are not always clear, even to the person making the decision. Producing a model for predictors/classifiers potentially exposes hidden motivations underlying a decision. This idea has been a particular feature of marketing research as a way of identifying and (and then exploiting) "gut reactions". Exposing such motivations in the context of peer reviewing would help reviewers and editors to increase awareness in and transparency of the peer review process, and this may again help to identify possible biases in decision-making.

Bias. 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 a tool based on machine learning is 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 the automated reviews may not adequately 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 among reviewers. The USA dominates the contribution to peer review, with 32.9% of all reviews vs. 25.4% of published article output (Publons, 2018). We suggest that AI systems can be used to expose possible biases and to inform actions taken to prevent their replication in future use of automated tools.

The ethics of (semi) automated peer review. As shown in section "Results", overfitting and other issues with the model we have used could lead to unintended consequences, like the creation of biased rules that could penalise under-represented groups or even

individuals if a tool such as the one we have developed is used in peer review. Following the categories delineated by Mittelstadt et al. (2016) and considering the most relevant of them to our study, we can identify three key examples of ethical concerns arising from our work:

Inscrutable evidence leading to opacity. When the link between the original data and the way they affect the model prediction is not easy to interpret, there is a problem of algorithm opacity, that can in turn lead to mistrust towards the algorithm and the data processors. An author will not trust a review if there is no transparency on the rationale for the decision taken. If tools are used to assist in decision-making of the sort we have described in future, it is crucial that there is as great a level of transparency as possible about how the models work to explain and justify decisions made.

Misguided evidence leading to bias. 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" (Macnish, 2012). Moreover, models based on machine learning, like the one described in this work, rely on past results (in this case, past reviews), and thus a model may propagate cultural and organisational biases already present in the learning set (Diakopoulos, 2016). Other sources of bias can be technical constraints or emergent contexts of usage. In review systems, a tool such as the one we have developed could in practice adversely affect decisions on papers produced by authors from low-income countries and or those on innovative topics if used without taking such possibilities into account.

Transformative effects leading to challenges for autonomy. Even using such models only to signal problematic papers or to assist reviewers could affect the agency of reviewers by creating new forms of understanding and conceptualisation. This may result in a specular effect to the one discussed in the previous point: the way the model interprets the manuscript could propagate to the reviewer, potentially creating an unintended biased outcome. For example, should the model identify as potential issues the presence of typos, the presence of references to papers from under-represented regions, or the usage of techniques that have been associated with previously rejected papers, the potential effect of the signalling of such issues to the reviewers could be an increase of the importance of such factors in the mind of the reviewers and influence their authority bias/status quo bias.

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.

Conclusions and future work

In this paper, we have reported an experiment involving three peer-reviewed conference proceedings, training a machine-learning system to infer a set of rules able to match the peer review outcome, ultimately providing an acceptance probability for other manuscripts. We focused on a rather superficial set of features of the submitted manuscripts, like word distribution, readability scores and document format.

Nevertheless, the machine-learning system was often able to successfully predict the peer review outcome: we found a strong correlation between such superficial features and the outcome of the review process as a whole.

We have seen how tools could be developed based on such systems, which could be used to create greater efficiency in the quality control and peer review process. We have also seen how

such tools could be used to gain insight on the reviewing process: our results suggest that such tools can create measurable benefits for scientometric studies, because of their explainability capability.

While the application of such AI tools is still in its infancy, we have observed some of the possible implications in terms of biases and ethics. Our findings point in the direction of a new type of analysis of typical human process, conducted with the help of machine-learning systems, one which is cognisant of ethical dimensions of the work as well as technical capabilities.

The following future work is suggested.

Feedback loop. We are interested in exploring the behaviour of reviewers when using these AI-powered support tools. We intend in future to carry out controlled experiments with academic reviewers, to understand the biases introduced by the AI signals on the reviewers. As discussed in "The ethics of (semi) automated peer review", understanding potential effects on the reviewers is fundamental to ensuring ethical usage of such tools.

Review process. When using openreview.net, we would be interested in taking into account the full text of the review itself (rather than only the review outcome) to better train the AI tools. A great deal of useful information is contained in the text of the reviews and rebuttals that inform the final decision.

Perception. Work on the first-impression bias needs to be extended, including more complex typographic layout indicators. Similarly, a more detailed analysis of the model could expose additional decision rules like language issues and formatting issues.

Disciplinary variation. We would like to explore how the design and application of AI tools carrying out semi-automated quality assessment can take place in the context of different disciplines, taking into account different disciplinary norms of communicating research results.

Grant applications. Funders might use such decision support systems to assess grant applications. Grant applications have a different structure (as they are proposing projects not reporting them), thus the content heterogeneity might be higher. We plan to investigate further the application of the methods discussed here to that domain.

Data availability

The datasets generated during the current study are not publicly available due to de-anonymisation risks, but are available from the corresponding author on reasonable request. The OpenReview data are available on https://openreview.net/.

Notes

1 Openreview is available at https://openreview.net/about.
2 Tokenisation is a standard machine-learning process, which consists in chopping a text into words (tokens), throwing away punctuation.
3 In statistics, F1-score is a measure of a test's accuracy, which considers both the precision and the recall.

References

AJE (2018) Peer review: how we found 15 million hours of lost time. URL https://www.aje.com/en/arc/peer-review-process-15-million-hours-lost-time, Accessed 20 Dec 2019

Akst J (2016) Researchers to CIHR: reverse peer review changes. URL https://www.the-scientist.com/the-nutshell/researchers-to-cihr-reverse-peer-review-changes-33236.

Ashley KD, Goldin IM (2011) Toward AI-enhanced computer-supported peer review in legal education. In: Biswas G, Bull S, Kay J, Mitrovic A (eds) JURIX. pp. 3–12

Balfour SP (2013) Assessing writing in MOOCs: automated essay scoring and calibrated peer review. Res Pract Assess 8:40–48

Barocas S, Selbst AD (2016) Big data's disparate impact. Cal Law Rev 104:671

Bornmann L (2011) Scientific peer review. Ann Rev Inform Sci Technol 45:197–245

Colman AM (1982) Manuscript evaluation by journal referees and editors: randomness or bias? Behav Brain Sci 5:205–206

Cozza V, Petrocchi M, Spognardi A (2016) A matter of words: NLP for quality evaluation of Wikipedia medical articles. In: Bozzon A, Cudré-Maroux P, Pautasso C (eds) International Conference on Web Engineering. Springer, pp. 448–456

Crossley SA, Allen DB, McNamara DS (2011) Text readability and intuitive simplification: a comparison of readability formulas. Read Foreign Lang 23:84–101

Crossley SA, Greenfield J, McNamara DS (2008) Assessing text readability using cognitively based indices. Tesol Quart 42:475–493

Cyranoski D (2019) Artificial intelligence is selecting grant reviewers in China. URL https://www.nature.com/articles/d41586-019-01517-8, Accessed 20 Dec 2019

D'Andrea R, O'Dwyer JP (2017) Can editors save peer review from peer reviewers? PLoS ONE 12:e0186111

Diakopoulos N (2016) Accountability in algorithmic decision making. Commun ACM 59:56–62

Florea L et al. (2019) From first impressions to selection decisions: the role of dispositional cognitive motivations in the employment interview. Person Rev 48:249–272

Ford E (2013) Defining and characterizing open peer review: a review of the literature. J Scholar Publish 44:311–326

Garcia M (2016) Racist in the machine: the disturbing implications of algorithmic bias. World Policy J 33:111–117

Goldin IM, Ashley KD (2011) Peering inside peer review with bayesian models. In: Biswas G, Bull S, Kay J and Mitrovic A (eds) International Conference on Artificial Intelligence in Education. Springer, pp. 90–97

Güçlütürk Y et al. (2017) Multimodal first impression analysis with deep residual networks. IEEE Trans Affect Comput 9:316–329

Harnad S (1999) Free at last: the future of peer-reviewed journals. D-Lib Magaz 5:12

Harnad S (1998) The invisible hand of peer review. Nature 5. https://doi.org/10.1038/nature28029.

Hojat M, Gonnela JS, Caelleigh AS (2003) Impartial judgment by the "gatekeepers" of science: fallibility and accountability in the peer review process. Adv Health Sci Educ 8:75–96

Horbach SPJM, Halffman W (2019) The ability of different peer review procedures to flag problematic publications. Scientometrics 118:339–373

Lee CJ et al. (2013) Bias in peer review. J Am Soc Inform Sci Technol 64:2–17

Lindgaard G et al. (2006) Attention web designers: you have 50 milliseconds to make a good first impression! Behav Inform Technol 25:115–126

Macnish K (2012) Unblinking eyes: the ethics of automating surveillance. Ethics Inform Technol 14:151–167

Mittelstadt BD et al. (2016) The ethics of algorithms: mapping the debate. Big Data Soc 3:68

Moys JL (2014) Typographic layout and first impressions: testing how changes in text layout influence reader's judgments of documents. Vis Lang 48(1): 881

Mulligan A, Hall L, Raphael E (2013) Peer review in a changing world: an international study measuring the attitudes of researchers. J Am Socr Inform Sci Technol 64:132–161

Nicholas D et al. (2015) Peer review: still king in the digital age. Learn Publ 28:15–21

Nuijten MB, Van Assen MALA, Hartgerink CHJ, Epskamp S, Wicherts JM et al. (2017) The validity of the tool "statcheck" in discovering statistical reporting inconsistencies. Preprint retrieved from https://psyarxiv.com/tcxaj/

Piech C, Huang J, Chen Z et al. (2013) Tuned models of peer assessment in MOOCs. In: D'Mello SK, Calvo RA and Olney A (eds) 6th International Conference on Educational Data Mining (EDM 2013). International Educational Data Mining Society, pp. 153–160

Pontille D, Torny D (2015) From manuscript evaluation to article valuation: the changing technologies of journal peer review. Human Stud 38:57–79

Preim J (2013) Beyond the paper. Nature 495:437–440

Publons (2018) Global state of peer review 2018. URL https://publons.com/static/Publons-Global-State-Of-Peer-Review-2018.pdf, Accessed 20 Dec 2019.

Ribeiro MT, Singh S, Guestrin, C (2016) Why should I trust you?: Explaining the predictions of any classifier. In: Balaji K, Mohak S (eds) Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, pp. 1135–1144

Ross-Hellauer T (2017) What is open peer review? A systematic review. F1000Res 6.588. https://doi.org/10.12688/f1000research.11369.2

Russell SJ, Norvig P (2016) Artificial intelligence: a modern approach. Pearson Education Limited, Malaysia

Schmidt T (2017) Statcheck does not work: All the numbers. Reply to Nuijten et al. (2017). PsyArXiv. http://psyarxiv.com/hr6qy.

Shanahan D (2016) A peerless review? Automating methodological and statistical review. Springer Nature BioMed Central, Research in progress blog. Available at: https://blogs.biomedcentral.com/bmcblog/2016/05/23/peerless-review-automating-methodological-statistical-review Accessed 6 Jan 2020

Siler K, Lee K, Bero L (2015) Measuring the effectiveness of scientific gatekeeping. Proc Natl Acad Sci 112:360–365

Sonntag D (2004) Assessing the quality of natural language text data. In: Dadam P, Reichert M (eds) GI Jahrestagung. pp. 259–263

Spezi V et al. (2018) Let the community decide? The vision and reality of soundness-only peer review in open-access mega-journals. J Document 74:137–161

SWEOR (2019) 27 eye-opening website statistics: is your website costing you clients? URL https://www.sweor.com/firstimpressions, Accessed 20 Dec 2019

Tennant JP (2018) The state of the art in peer review. FEMS Microbiol Lett 365 (19). https://doi.org/10.1093/femsle/fny204.

Tennant JP, Dugan JM, Graziotin D et al. (2017) A multi-disciplinary perspective on emergent and future innovations in peer review. F1000Res 6:1151. https://doi.org/10.12688/f1000research.12037.3

Wood TJ et al. (2018) Can physician examiners overcome their first impression when examinee performance changes? Adv Health Sci Educ 23:721–732

Zarsky T (2016) The trouble with algorithmic decisions: an analytic road map to examine efficiency and fairness in automated and opaque decision making. Sci Technol Human Value 41:118–132

Competing interests

The authors declare no competing interests.

Additional information