Technology assisted research assessment: Algorithmic bias and transparency issues[1]

Mike Thelwall and Kayvan Kousha

Abstract

Purpose: Technology is sometimes used to support assessments of academic research in the form of automatically generated bibliometrics for reviewers to consult during their evaluations or by replacing some or all human judgements. With Artificial Intelligence (AI), there is increasing scope to use technology to assist research assessment processes in new ways. Since transparency and fairness are widely considered important for research assessment and AI introduces new issues, this review investigates their implications.

Design/methodology/approach: This articles reviews and briefly summarises transparency and fairness concerns in general terms and through the issues that they raise for various types of Technology Assisted Research Assessment (TARA).

Findings: Whilst TARA can have varying levels of problems with both transparency and bias, in most contexts it is unclear whether it worsens the transparency and bias problems that are inherent in peer review.

Originality: This is the first analysis that focuses on algorithmic bias and transparency issues for technology assisted research assessment.

Keywords: Technology Assisted Research Assessment; bibliometrics; research evaluation; machine learning, algorithmic bias, transparency

1 Introduction

Technology Assisted Research Assessment (TARA) refers to the use of routine computer automation or artificial intelligence to generate information to support or replace human judgement for research evaluations. TARA may have value if it improves outcomes or saves the time of administrators or skilled researchers without introducing perverse incentives into the systems (Wilsdon et al., 2015). TARA has previously taken the form of mostly hidden computerisation of bibliometric databases and bibliometric indicator calculations (e.g., article citation counts, Journal Impact Factors) but with the rise of Artificial Intelligence (AI), it has become technologically possible to provide a wider range of functionalities, such as identifying or selecting evaluators (Fiez et al., 2020), detecting plagiarism (Zhang, 2010), checking methods details (Wren, 2018), and estimating the overall quality of articles from bibliometrics and/or text and/or other metadata (Thelwall et al., 2023).

Both transparency and bias are important concerns for research assessment (Hicks et al., 2015). Whilst bias against researchers or institutions is undesirable from a natural human justice perspective, bias against aspects of research, such as fields, methods, or output types, is undesirable from a systemic perspective if it provides perverse incentives to researchers to alter their behaviours in ways that do not benefit science, such as by changing to a higher citation field. Transparency is also important from a natural justice perspective to allow those evaluated to check the key assumptions and calculations in TARA data and, if necessary, challenge errors or inappropriate calculations.

[1] Thelwall, M. & Kousha, K. (in press). Technology assisted research assessment: Algorithmic bias and transparency issues. Aslib Journal of Information Management. https://doi.org/10.1108/AJIM-04-2023-0119

It is already recognised that bibliometric data often requires extensive computer processing, and potentially reducing transparency and generating biases (Hicks et al., 2015). This review extends previous discussions of bibliometric transparency and bias to the wider context of TARA, incorporating insights from analyses of general AI transparency and bias issues.

2 Transparency in technology assisted research assessment

This section covers transparency in the sense of the ability of the people assessed to fully understand the procedures used to assess them. Whilst reproducibility is also important for research assessment, a fully reproducible complex assessment may not be transparent to the person assessed. As mentioned above, transparency in assessment allows those assessed to check the results and suggest corrections for mistakes, when necessary. One of the ten principles of the influential Leiden Manifesto for research evaluation is, "Keep data collection and analytical processes open, transparent and simple" (Hicks et al., 2015). In practice, simplicity is a core aspect of transparency because those evaluated may not be able to understand complex or large-scale computing solutions even if they are fully public. Unfortunately, since research is complex and carried out on a large scale internationally, all TARA probably has either obvious or hidden complexity, as described below, limiting its transparency. In situations where adequate transparency is impossible because complexity is necessary for sufficient accuracy, reproducibility might sometimes be judged to be an acceptable alternative.

2.1 Publication databases

2.1.1 Coverage

The publication databases used in research evaluation are mostly controlled by commercial organisations such as Dimensions.ai (Digital Science), Scopus (Elsevier) and the Web of Science (Clarivate). These organisations broadly publish their methodologies for finding and including journal articles. The main sources are manually curated lists of academic journals for Scopus[2] and the Web of Science[3], which are published and public. The process of choosing these journals is human-based and private, although the outcome is public.

Elsevier and Clarivate presumably have agreements with the publishers to harvest relevant information about the journals from the publishers' websites and then use their own private algorithms to transform the raw data into bibliometric information, by extracting titles, matching articles with journals and metadata such as DOIs, author names and affiliations, assuming that they are not already supplied in XML or other form by the publisher. This basic processing seems uncontroversial, even if the precise algorithms are not used. Nevertheless, there can be errors for unusual cases, such as for conference papers dual published as journal articles and for consortia listed as authors of academic papers.

Dimensions.ai uses public Crossref data provided freely by publishers as well as arrangements with other publishers to directly harvest their bibliometric metadata, and

[2] https://www.elsevier.com/solutions/scopus/how-scopus-works/content
[3] https://mjl.clarivate.com/search-results

crawlers to harvest various repositories, such as PubMed and arXiv[4]. It does not publish a list of journals indexed but explains how to check if a journal is indexed[5].

2.1.2 Metadata and citation indexing

In addition to ingesting publication information, bibliometric databases typically have layers of extra information from largely or completely automated processes to add value to end users. These algorithms range from simple heuristics to deal with routine or unexpected information to AI for non-trivial tasks. Routine automation tasks include the following.

- Identifying and dealing with apparent data errors, such as malformed DOIs.
- Matching affiliations to author names for an article.
- Connecting author names in an article to author IDs to connect all the articles by the same author and support searches by researcher rather than by article.
- Identifying multiple copies of the same publication from different sources.
- Matching reference lists to cited documents in the absence of DOIs.

The last of these is the most important for research evaluation since errors in reference matching can reduce citation counts (e.g., Harzing, 2017; van Eck & Waltman, 2019), and duplicate publications can cause the same problem by sharing citations (van Eck & Waltman, 2019). Errors can originate from many minor sources, and no commercial organisation seems to have published the algorithms used, which may be regarded as commercial secrets. These algorithms presumably use a range of heuristics to cope with slightly different versions of article titles and author names, and perhaps also more sophisticated AI to connect citations to preprints with citations to the published version. In practice, the complexity of such processes probably means that publishing transparent versions would not be very helpful. Citation indexing errors should become rarer over time with the inclusion of DOIs within reference lists.

2.1.3 Field classification

Classifying articles into academic fields is important in research evaluation because the fields are used to identify the reference set for each article for research indicator calculations or evaluations. In Scopus and the Web of Science, journals are first manually classified into one or more fields, then all articles in each journal are assigned to all the fields of that journal. In this context, the procedure used to classify journals into fields categories is the key to transparency. For both organisations, the classification seems to be manual but probably helped by automated analyses of the references and citations of each journal to see which other journals and fields it connects to most frequently. Journal classification is an important aspect of non-transparency because a journal's categories can have a substantial influence on whether its articles tend to be cited above or below the world average for its categories.

The most accurate taxonomies of science algorithms that operate at the article level rather than the journal level (Klavans & Boyack, 2017). Dimensions.ai uses an AI algorithm to classify articles into fields rather than using a primary journal-based classification system (Hook et al., 2018). Clarivate also has the option to classify articles by field algorithmically for field normalised indicator calculations for its InCites tool (the Citation Topics option at: https://incites.help.clarivate.com/Content/Research-Areas/research-areas.htm). In the latter case, the algorithm is fully public, but it is too complex to be fully transparent to end

[4] https://plus.dimensions.ai/support/solutions/articles/23000018860-how-is-the-publications-data-harvested-
[5] https://www.dimensions.ai/submit-journal-and-book-titles/

users because it depends on relationships between tens of millions of references and/or keywords (see also: Ruiz-Castillo & Waltman, 2015). This is arguably a minor issue since the human decision making of Clarivate and Elsevier for journal-based classification is similarly opaque.

2.1.4 Research indicator calculations

Bibliometric databases report a range of article-level and journal-level indicators, often in products that are available separately from the main database (e.g., InCites from Clarivate Analytics, Perspectives & Insights from Dimensions.ai). These include journal impact indicators, such as the Journal Impact Factor (JIF) and field normalised citation counts or percentiles for papers or sets of papers (e.g., paper x is in the top 10% cited for its field).

Some of the indicators reported by bibliometric databases seem to be relatively transparent in the sense that the formulae are published and tend to be simple and checkable. This strategy presumably helps scientists to understand and adopt them. Clarivate[6], Elsevier[7] and Dimensions[8] publish formulae for their indicators and explain them. The JIF is an example of a relatively transparent calculation because it is the simple ratio of the number of citations from items published in a given year to items in a journal published in the previous two years, divided by the number of articles published in that journal in the previous two years. This still contains elements of hidden transparency, however, in the form of the selection of journals to index in the first place (which affects the JIF numerator) and the AI to match cited and citing items. Unfortunately, the better algorithms tend to be less transparent. For example, field normalised journal impact indicators are an improvement on the JIF because they take into account that citation rates vary naturally between academic fields, but field normalised citation impact formulae are more complex and have added complexity through whichever field classification procedure is used.

Some citation indicators, such as SCImago Journal Rank, are too complex to be transparent, even if fully described (Mañana-Rodríguez, 2015), because they rely on large matrix factorisations that integrate the entire bibliometric database in one high-dimensional matrix calculation that a human could not understand. Another non-transparent indicator is the Relative Citation Ratio (Hutchins et al., 2016) because insufficient information is published to make it reproducible.

2.2 Peer review databases

AI software has been developed to write or evaluate peer review assessments of academic journal articles (e.g., Kumar et al., 2022). If this software is used, then the datasets used to develop it create an indirect transparency issue. Since machine learning AI software all works by identifying patterns, its capabilities depend on the data it is trained on. Essentially, the software will try to replicate is training data. Thus, peer review software developed on one type of data may not work well on another, creating an accuracy problem and, indirectly, a transparency issue. As an extreme example, peer review software developed exclusively on physics articles and/or peer review reports may be unable to identify important ethical concerns when it is applied to health-related research. Thus, a transparency issue if/when peer review AI software is applied in research evaluation is that those evaluated need to know

[6] https://clarivate.com/webofsciencegroup/wp-content/uploads/sites/2/2021/06/JCR_2021_Reference_Guide.pdf
[7] https://www.elsevier.com/solutions/scopus/how-scopus-works/metrics/citescore
[8] https://plus.dimensions.ai/support/solutions/folders/23000031268

what type of data the software was trained on. This is also an efficiency and accuracy consideration.

Most peer review reports are not published and so not available for training machine learning solutions. The main exceptions include many BioMed Central journals, some articles in most MDPI journals, and a few individual journals including the BMJ, Quantitative Science Studies, SciPost Physics, and the F1000 publication platform. Publishers producing in-house peer review AI may also be able to use their own private peer review reports. Software trained on in-house private peer review would clearly have a transparency issue.

2.3 AI algorithms for TARA

Many TARA tasks benefit from AI algorithms. As mentioned above, these include identifying evaluators, detecting plagiarism, checking methods details, and estimating the overall quality or future citation impact of articles from bibliometrics and/or text and/or other metadata. This section discusses the extent to which AI algorithms for these tasks can be transparent. For example, thresholding, regression or AI approaches have been proposed to estimate the quality of academic publications or their future citation counts from bibliometric information (e.g., Chen & Zhang, 2015; HEFCE, 2015; Thelwall et al., 2022; Traag & Waltman, 2019). If such a system is used in important applications, then end user understanding and transparency may help to give confidence in the system and allow researchers to verify the input and understand all the steps that the algorithm used to get the answer (e.g., a score or recommended assessor for an output). This would typically sacrifice accuracy, however, since state of the art algorithms are not transparent for most machine learning tasks (see below).

2.3.1 Opaque AI algorithms

Away from the field of research evaluation, AI researchers have discussed the problem that most machine learning algorithms are almost completely opaque in the sense of being too complex for an intuitive understanding of how they work in a particular case.

Deep learning is an example of an effective but opaque AI approach. A deep learning model may be a neural network with thousands of interconnected nodes, with each connection having its own weights. Whilst the input and output layers may be interpretable, the intermediate layers may not have an intuitive understanding even if there were not too many nodes to follow. This is even more complex that the matrix case mentioned above because the matrices used for bibliometric indicators are two dimensional (n x n, where n is the number of journals, for example) and matrix calculations are typically simple, whereas neural networks can have many more than two layers (n x m x …) and connections between layers can be driven by functions rather than simple formulae. Support Vector Machine (SVM) algorithms are opaque through complexity, not allowing an intuitive understanding of how they work for a specific problem. This is because SVMs operate in high dimensional spaces that are beyond human understanding.

For contrast, the decision tree is an example of a relatively non-opaque simple algorithm that is easy to understand because it requires checking multiple transparent decisions. A decision tree has a set of binary (or n-way) decisions that lead to the final recommendations. Figure 1 illustrates a decision tree that might be made to decide whether a cited article matches a given article in the database in the absence of a DOI. Ignoring the term “nearly” in the tree, which could be operationalised with a precise formula, this is fully transparent in the sense that it is easy to identify why a match was made or not. Nevertheless, decision trees still have a degree of opaqueness in the sense that they are built by complex

AI algorithms that made decisions about which nodes to add by identifying patterns within the (huge) database.

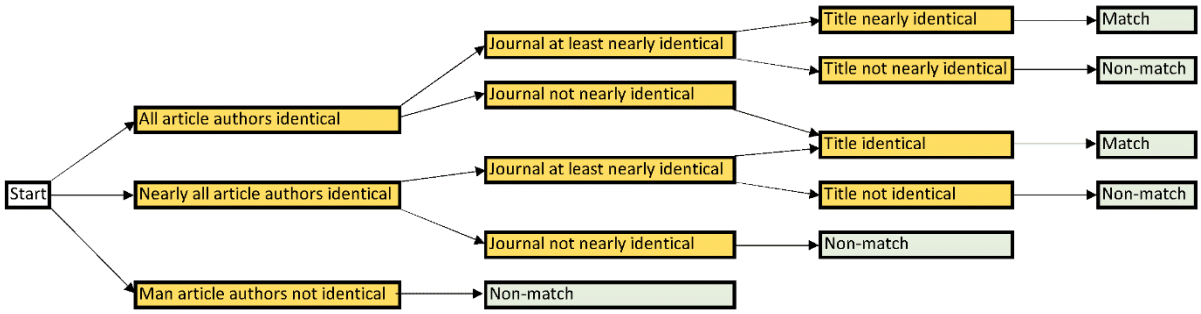

Figure 1. A decision tree to decide whether a cited reference matches an article in a bibliometric database.

Three current state-of-the-art machine learning algorithms, random forest, gradient based classifier, and extreme gradient boost, all use hundreds of simultaneous decision trees, combining them using mathematical formulae for the output (Chen et al., 2015). In this case, all their building blocks are just transparent decision trees, the algorithms overall are opaque due to their complexity because it would be unreasonable to check 200 decision trees for each decision made, and impossible to intuitively understand the effect of combining 200 decision trees.

2.3.2 Explainable AI

Algorithmic opaqueness makes it more difficult to check that an algorithm has not introduced made biases and makes it more difficult for the algorithm owner to be accountable for the decisions (e.g., Diakopoulos & Koliska, 2017). This has led to the field of eXplainable AI (XAI) or "white box" AI (Vilone & Longo, 2020; Xu et al., 2019), which focuses on algorithms with decision making process that a human expert could understand, such as linear regression, a finite set of rules, or a decision tree (as above). This might also allow a specialist to adjust part of the AI based on their knowledge that it was incorrect even though it was consistent with the dataset that the AI had been trained on (Gunning et al., 2019). There are different grades of transparency in XAI, with the most transparent being explainable to end users rather than AI experts.

As mentioned above for decision trees, even XAI built on large datasets is not fully transparent in the building process because it relies on identifying patterns in huge amounts of data, even if the final result is understandable.

2.3.3 Implications for TARA and comparison with human judgements

As the above abstract discussion shows, most machine learning algorithms, including the most accurate, are opaque. Thus, the most accurate systems to estimate article quality, predict long term citation counts, or support any other complex research assessment task are likely to be fully opaque. Whilst the opaqueness of AI is concerning, the same is true of aspects of peer review, as discussed below.

3 Peer review transparency

Although peer review is the has been the cornerstone of academic quality judgement for at least half a century, it varies from being relatively transparent fully opaque. Many types of peer review have transparency in the judgements in the form of an explicit justification for a score or recommendation. For example, a journal or conference would normally return a set of peer review reports to an author alongside the editor's decision (e.g., publish, major revisions), and the same is true for grant proposals for many funders. Similarly, a rejected candidate at an academic interview might expect to be given feedback about why they were not selected. In other contexts, no feedback might be given, such as if a candidate is not selected for an interview or an academic prize.

One of the most transparent forms of academic peer review is fully open journal peer review, where the authors and reviewers know each other's identities and the reviewer report and recommendation is published online. Here, the decision, decision maker and main reasons are transparent. Nevertheless, there are still two important aspects that are typically obscured. First, journal article reviewers are normally selected by journal editors using their expertise and/or technology (e.g., keyword searches or text matching suggestions from a publication management system) and the details are usually not fully explained. In any case, the expert knowledge of the selecting editor is tacit and so it would be impossible to write it down in the form of an algorithm, although the editor might be able to write a justification of their choice. The same is true for reviewer reports: although a reviewer can describe shortcomings in a paper and write a justification for their decision, the outcome is necessarily derived at least in part from their tacit knowledge and subjective judgements so the process leading to the peer review report and overall recommendation is opaque, even if it can be justified by the report. Of course, most journal peer review anonymises reviewer identities and this is believed to increase its credibility (e.g., Karhulahti & Backe, 2021).

At the other extreme, an example of fully opaque peer review is the national research assessment system of the UK, the UK Research Excellence Framework (REF, www.ref.ac.uk). In particular, the REF panel members (1000+ mostly senior researchers) that conduce the expert review use their subject expertise, knowledge of the REF rules and discussions with other panel members to reach decisions about the quality score to allocate to each output. Many of the decisions about scores are probably based on intuition with a component of emotional reaction, "is this research exciting", rather than through simple explainable processes, especially in the arts and humanities. In any case, the decision-making process is not communicated to the output authors and so is 100% opaque. Instead, authors are only given vague feedback about large sets of outputs, such as "within the [set of 100 outputs submitted] those on the topic of [x] were considered particularly strong".

As the above examples illustrate, non-transparent forms of TARA would not necessarily result in a loss of transparency for those assessed because peer review always has important non-transparent components including the cognitive processes leading to scoring decisions or recommendations. Thus, forms of non-opaque TARA can potentially increase transparency.

4 Bias in technology assisted research assessment

Bias is an "inclination or prejudice for or against one person or group, especially in a way considered to be unfair"[9]. In the context of research evaluations, biases might be against

[9] https://www.lexico.com/definition/bias

individual people, institutions, research methods, genders, career stages, output types, or negative findings.

4.1 Bibliometric biases

When bibliometric data is used to support assessment then there is the potential to introduce many types of bias, and some are summarised here. Although gender bias is widely believed to occur (e.g., Rowson et al., 2021), this is not an issue from and article-level evaluation perspective because female first-authored articles tend to be slightly more cited than male first-authored articles in the UK (Thelwall, 2020). This is counter-intuitive because men typically dominate citation-based lists based on career citations or the h-index. This domination tends to happen because men tend to have fewer career gaps, are less likely to leave academia, and retire later. Because of these factors, they tend to accrue more career citations. In addition, today's older academics started when there were larger obstacles to women entering academia than there are today.

Field biases: Academic fields cite at different rates, with different length reference lists, citing different balances of journal articles, books, and other outputs, and citing different age outputs. Because of this, average citation counts differ substantially between fields. Citation counts should therefore only be compared between articles from the same field, unless field normalised or percentile indicators are used instead (Bornmann et al., 2013; Thelwall, 2017; Waltman, et al., 2011). This also applies to Journal Impact Factors, which should not be compared between fields. Of course, citation counts should also not be compared between articles of different ages, unless with field normalised or percentile scores.

Research type biases: Some types of research are naturally more cited than others, which introduces another citation bias. Review papers are the clearest example of a type that is usually more cited (Aksnes, 2003). Articles using some methods can also tend to be more highly cited (Antonakis et al., 2014; Fairclough & Thelwall, 2022; Thelwall & Nevill, 2021). In particular, it seems likely that, within mixed methods fields, papers making more hierarchical contributions (e.g., incremental method improvements) or contributing to faster publishing specialisms (e.g., simulation modelling rather than interview-based studies) will tend to be more cited, or at least cited more quickly. Papers in an expanding research area are also likely to be more cited because there are relatively many citing papers compared to the number of potentially cited papers. Positive results are also more likely to be cited (e.g., Jannot et al., 2013; Tincani & Travers, 2019; Urlings et al., 2021), although in REF terms these might also be judged to be more significant.

Country biases: Citation bias is likely against research from, about, or in the languages of, countries that are not well indexed in the bibliometric databases used for a citation analysis. All major citation databases make decisions about which journals to cover, and, as mentioned above, they seem to primarily cater for English-language documents so this leads to a bias against research that is from countries where research is often not written in English (Mongeon & Paul-Hus, 2016; van Leeuwen et al., 2001; Vera-Baceta et al., 2019). Since researchers are disproportionately cited from their own country (Lancho Barrantes, et al., 2012; Thelwall & Maflahi, 2015), under-indexing the work of a country creates a citation bias against the few articles from that country that are indexed. This is exacerbated for nationally-focused research that would expect to rarely be cited from other countries, perhaps including studies on national politics, indigenous plants and animals. Even for countries that are well indexed by all major databases, academics with interests that focus on less well-indexed

countries (e.g., some Area Studies) may be disadvantaged in national evaluations. Another issue is that research evaluators may subscribe to part of a bibliometric database, such as the core Web of Science without the Chinese Science Citation Database component, because of cost considerations, causing bias against research mainly indexed in the omitted sections.

Research volume bias: Related to country biases, an article on a topic that few researchers are publishing about may tend to be less cited than articles on popular topics. This may be legitimate if the more researched topic is more important but not legitimate if the topics are equally important but there is more activity about one topic for economic reasons. For example, an ecological researcher in a region with a relatively unusual characteristics and few researchers may be rarely cited for this reason (Culumber et al., 2019).

Recognition/prestige bias: Researchers may prefer to cite work from well-known people (the "Matthew effect", Merton, 1968), or prestigious sources (journals, institutions) because they are biased in its favour or consider it to be a safe option. Well known works can also be cited as concept markers for a topic rather than for their contents (Case & Higgins, 2000).

4.2 Algorithmic bias

Algorithms can show bias and make biased decisions (Kordzadeh & Ghasemaghaei, 2021; Mehrabi al., 2021; Navarro et al., 2021), as illustrated by some high-profile cases outside of scientific applications. For example, a recruiting tool from Amazon was discontinued after it was shown to be biased against women[10]. AI systems can be biased because they are fed biased rules, learn from biased data, or accidentally introduce bias as a side-effect of something else. There are different types of algorithmic bias.

Design bias: This can occur if a system is poorly designed. For example, a facial recognition system that is only trained on white faces because of the prejudice or thoughtlessness of its creators would be biased (Furl et al., 2002; Lee, 2018) and this could have unpleasant effects when it is used in practice. Alternatively, an inappropriate set of inputs to a system might be selected so that it is not shown important information because the designers did not realise its value. For example, an AI system to estimate the quality of candidates based on their career achievements would be biased against women if it was not fed career gap information.

Existing bias: The system learns existing prejudices in society from its input data and conforms to them. For example, since some job categories are heavily gendered (e.g., nurse, carpenter), a machine learning system designed to recommend jobs to candidates based on their CVs could easily learn and then exacerbate existing gender divisions by only recommending carpentry to men and nursing to women. Such an algorithm might also primarily recommend senior jobs to men, or lower paid jobs to ethnic minority candidates. Here the system notices a pattern (e.g., most previously interviewed candidates for top jobs have been male) and then uses the gender on a CV, together with other information, to help predict whether the person should apply for a senior role. Whilst women and nonbinary people might still be recommended to apply, their CVs would have to be better to trigger this recommendation.

Indirect bias: An AI system makes biased decisions because of factors unrelated to its primary design goals. For example, a system paying to show adverts to users of an electronic

[10] https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

system might primarily target the cheapest demographic to reach the largest audience. This might lead to career adverts disproportionately targeting the cheapest gender (Lambrecht & Tucker, 2019) or age group unless the system is configured specifically for demographic equality. Similarly, sentiment analysis systems have been shown to disproportionately reflect the opinions of demographics that express sentiment most clearly, such as women compared to men (Thelwall, 2018).

4.3 Bias in AI-based TARA

A TARA system that uses machine learning can reflect existing biases in the data use to build it or can generate new types of bias, as outlined above. This section reviews the practical implications for various research assessment tasks and applications.

A system to predict peer review scores by learning patterns associated with research quality from bibliometric and other data is likely to inherit some but not all the biases of bibliometrics and peer review and may even generate new ones, as discussed above. In terms of bibliometric inputs, higher citation counts associate with higher quality research to varying extents in most fields, so an AI system is likely to leverage citation counts. If it is fed field normalised citation counts rather than raw citation counts, then this will avoid substantial biases against low citation fields. Even with field normalised specialisms, AI systems will still inherit biases against low citation types of research, as well as the country, prestige and research volume biases discussed above. Since the AI system will learn from peer review scores and assuming that the peer review scores did not reflect the same biases as the citations, then the AI system would, in theory, be able to learn to correct the AI bias with the human scores. In practice, this is unlikely to work perfectly because an AI system is unlikely to be fed with enough training data to learn any patterns reflected in a small minority of the article scores. Thus, depending on the volume of training data and the number of articles in the set that the bias is against, the bibliometric bias may be largely replicated by the AI or partially bypassed.

One previous study has developed an AI system to estimate the quality of research articles and evaluated whether it has biases, in comparison to peer review scores, along multiple dimensions. Using provisional REF2021 peer review scores and an AI system designed to predict them from bibliometric data, it found that the most accurate AI solution did not create (or correct) biases against women, early career researchers or larger institutions (which tend to be more prestigious in the UK) but it did create a minor bias against departments publishing higher quality research because the AI errors tended to reduce their higher scores (except in Chemistry) (Thelwall et al., 2022).

Some but not all peer review biases are also likely to be learned by an AI system that predicts quality scores and is trained on a set of journal articles with bibliometric information and peer review scores, or that learns to write peer review reports based on an existing collection of article and reviews. This essentially depends on whether the relevant biasing information is fed into the AI system in the learning phase and the variety of the reviewer judgements. In the case of prestige information, if the AI system is not fed author career information, then it could not directly learn a prestige bias from the human reviewer scores and if it is not fed the gender and nationality of the authors then it cannot learn gender and nationality bias directly, even if it is present in the peer review scores. AI is also likely to ignore layout bias, as it would presumably not be fed with layout information.

A system could learn cognitive distance bias and confirmation bias if they dominated the peer review scores of relevant articles. For example, if all education reviewers gave low

scores to qualitative research because they thought that quantitative research was inherently superior, then the AI system would probably learn to be biased against qualitative research. On the other hand, if the reviewers were evenly split between those that favoured quantitative and those that favoured qualitative research, then the AI system may well not learn a qual/quant bias and be less biased than individual reviewers in this regard. Similarly, if one topic was cognitively distant from all reviewers then the AI might learn to allocate lower scores to that topic.

Finally, research assessments many need to use linguistic TARA systems to support some of its tasks. Automatic translation systems can introduce gender biases (Prates et al., 2020) and so AI systems relying on translation (e.g., for articles not written in English and without an English translation) may introduce gender biases. AI systems processing textual input as part of quality score prediction may generate biases against minority groups through language expression (Cheuk, 2021). For example, one empirical study has developed AI systems to predict conference review accept/reject decisions from word frequency text analysis of the submitted papers. The factors found most useful by the system were all superficial and indirectly associated with higher quality rather than measures of it: avoiding "quadratic", few sentences, many difficult words, many pages, and many syllables per word (Checco et al., 2021). This approach seems likely to generate a bias against non-native English speakers who may prefer to use more straightforward language.

5 Peer review bias

Several factors are known to influence peer review decisions, as summarised in a recent comprehensive review (Lee et al., 2013). These biases might be removed or reduced by TARA, including bibliometrics, but they are also likely to translate into citation biases if academics have similar biases when deciding what to cite. It is known that even the most expert academic peer reviewers sometimes make poor decisions, such as editorial rejection of important articles (Siler et al., 2015), but this section focuses on systematically sub-optimal decisions with an identifiable cause.

Prestige bias: Reviewers may form more favourable judgements for outputs from successful researchers (Merton, 1968; Tran et al., 2020), from more prestigious institutions, or for articles that they believe are standard to cite in the field (Brooks, 1986).

Nepotism: Academic reviewers may form more favourable judgments of the work of people that they know (Sandström & Hällsten, 2008).

Gender bias: Although universities have historically been extremely sexist institutions, there is not a consensus about whether gender bias in academic evaluations remains a problem. There is not strong empirical evidence of overall gender bias in judgements (Ceci et al., 2011) despite persistent problems with the underrepresentation of women in senior positions. Nevertheless, there are areas or aspects of science that are chilly climates for female researchers (Biggs et al., 2018).

Nationality/ethnicity: Reviewers may be prejudiced against the work of academics from particular countries or ethnicities (Hojat et al., 2003).

Cognitive bias and distance: This occurs when judgments are influenced by the reviewers' beliefs about the subject matter without considering whether their beliefs are universal (e.g., Bader et al., 2021). This can occur in two ways: a researcher from a distant field may undervalue a study through a lack of understanding of its importance, or a researcher from a competitive paradigm may not value a study at all. Thus, variations of cognitive bias seem likely to be widespread or universal and unavoidable, at least in the first

form. Nevertheless, empirical evidence in limited contexts shows the opposite of what might be predicted: reviewers are stricter on topics closer to their own area (Boudreau et al., 2016; Wang & Sandström, 2015). Cognitive distance presumably applies to all interdisciplinary research to some extent, since reviewers seem more likely to be unfamiliar with some of the component disciplines (Rinia et al., 2001).

Confirmation bias: A reviewer may be more critical of work that challenges their beliefs (Mahoney, 1977).

Novelty bias: The most novel research can sometimes have difficulty in passing peer review and eventually be published in less prestigious journals than the subject merits (Campanario, 2009; Gans & Shepherd, 1994; Wang et al., 2017).

Layout bias: Reviewers may be influenced by first impressions based on article layout (e.g., Moys, 2014). For example, if they review a preprint in an awkward format (e.g., double spaced, with figures and tables at the end) they may be more likely to give a negative evaluation than if they had read the journal printed version.

As the many examples above suggest, there are many known potential biases in peer review. The presence and strength of these biases is likely to vary substantially between contexts, including by country, assessment purpose, and assessor experience. It is very difficult to detect them in practice because peer review is the gold standard for research assessment. Because of this it seems impossible to be sure whether TARA biases, if known, increase or decrease human biases, and are stronger or weaker than them.

6 Conclusions: Transparency, bias and perverse incentives

As the review above shows, TARA can lack transparency either through hidden processes or algorithms that are opaque by design or through complexity. In situations in which they replace or support more fully transparent evaluations, such as peer review where the reviewers must explicitly justify scores, this can be unfortunate. In situations where the peer review decision process are opaque to those reviewed, lack of transparency in TARA can still be a problem from the perspective of checking for bias (except in decisions) and more effectively supporting reviewers. Nevertheless, since peer review is never fully transparent and some aspects of TARA are usually not opaque, such as the inputs if not the algorithms, introducing TARA is likely to often increase some aspects of transparency in research assessments and decrease others. When deciding whether to use TARA it is therefore important to identify and evaluate both of these changes.

Using TARA to support or replace peer review in research assessment could introduce new biases or correct peer review biases (e.g., nepotism). There is little strong evidence about the net effect of bibliometrics, the most important current type of TARA, because there are field differences in its value and, since peer review also has biases, there are no ground truths to compare bibliometrics (or peer review) against. Nevertheless, the most comprehensive study so far suggests that bibliometrics if used in an AI system to estimate published journal article quality for a single country, would not increase or decrease gender, institutional size, or early career researcher biases compared to expert peer review (Thelwall et al., 2022). Of course, even if a fully automatic research evaluation system is used, human judgements will still play a role in the selection of the algorithms and input data as well as the interpretation of, and actions based on, the results. Thus, automatic solutions do not bypass the need for fair human judgements within the wider research evaluation system. Automation would also not bypass the need for understanding the limitations of citation counts as evidence in research evaluation and likely inputs into future technology assisted assessment systems.

A final very important consideration for research assessments is whether they generate perverse incentives. In this context, transparency is a *disadvantage* for research assessments where those assessed are told the assessment procedure in advance because this gives an opportunity to change behaviour towards the thing assessed. For example, in the Italian assessment system (Ancaiani et al., 2015), knowing that JIFs and citation counts will be used in some fields increases transparency but incentivises publishing in high impact journals. This may be seen as a perverse incentive since the researchers might otherwise choose a journal that gets a larger audience, such as a national journal for an issue of mainly national relevance. Peer review has an advantage here, assuming that academics will not try to persuade the reviewers, in that the cognitive processes they use are opaque. Of course, there is still some theoretical potential for perverse incentives, such as citing or befriending the assessors, but this seems to be minor in comparison.

7 Acknowledgements

This study was funded by Research England, Scottish Funding Council, Higher Education Funding Council for Wales, and Department for the Economy, Northern Ireland as part of the Future Research Assessment Programme (https://www.jisc.ac.uk/future-research-assessment-programme). The content is solely the responsibility of the authors and does not necessarily represent the official views of the funders.

8 References

Aksnes, D. W. (2003), "Characteristics of highly cited papers", Research Evaluation, Vol. 12 No. 3, pp. 159-170.

Ancaiani, A., Anfossi, A. F., Barbara, A., Benedetto, S., Blasi, B., Carletti, V., and Sileoni, S. (2015), "Evaluating scientific research in Italy: The 2004–10 research evaluation exercise", Research Evaluation, Vol. 24 No. 3, pp. 242-255.

Antonakis, J., Bastardoz, N., Liu, Y., and Schriesheim, C. A. (2014), "What makes articles highly cited?", The Leadership Quarterly, Vol. 25 No. 1, pp. 152-179.

Bader, H., Abdulelah, M., Maghnam, R., and Chin, D. (2021), "Clinical peer review; A mandatory process with potential inherent bias in desperate need of reform", Journal of Community Hospital Internal Medicine Perspectives, Vol. 11 No. 6, pp. 817-820.

Biggs, J., Hawley, P. H., and Biernat, M. (2018), "The academic conference as a chilly climate for women: Effects of gender representation on experiences of sexism, coping responses, and career intentions", Sex Roles, Vol. 78 No. 5, pp. 394-408.

Bornmann, L., Leydesdorff, L., and Mutz, R. (2013), "The use of percentiles and percentile rank classes in the analysis of bibliometric data: Opportunities and limits", Journal of Informetrics, Vol. 7 No. 1, pp. 158-165.

Boudreau, K. J., Guinan, E. C., Lakhani, K. R., and Riedl, C. (2016), "Looking across and looking beyond the knowledge frontier: Intellectual distance, novelty, and resource allocation in science", Management Science, Vol. 62 No. 10, pp. 2765-2783.

Brooks, T. A. (1986), "Evidence of complex citer motivations", Journal of the American Society for Information Science, Vol. 37 No. 1, pp. 34-36.

Campanario, J. (2009), "Rejecting and resisting Nobel class discoveries: accounts by Nobel Laureates", Scientometrics, Vol. 81 No. 2, pp. 549-565.

Case, D. O., and Higgins, G. M. (2000), "How can we investigate citation behavior? A study of reasons for citing literature in communication", Journal of the American Society for Information Science, Vol. 51 No. 7, pp. 635-645.

Ceci, S. J., and Williams, W. M. (2011), "Understanding current causes of women's underrepresentation in science", Proceedings of the National Academy of Sciences, Vol. 108 No. 8, pp. 3157-3162.

Checco, A., Bracciale, L., Loreti, P., Pinfield, S., and Bianchi, G. (2021), "AI-assisted peer review", Humanities and Social Sciences Communications, Vol. 8 No. 1, pp. 1-11.

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., and Chen, K. (2015), "Xgboost: extreme gradient boosting. R package version 0.4-2", https://xgboost.readthedocs.io/en/stable/

Chen, J., & Zhang, C. (2015), "Predicting citation counts of papers. In 2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI&CC)" (pp. 434-440), Los Alamitos: IEEE Press.

Cheuk, T. (2021), "Can AI be racist? Color-evasiveness in the application of machine learning to science assessments", Science Education, Vol. 105 No. 5, pp. 825-836.

Culumber, Z. W., Anaya-Rojas, J. M., Booker, W. W., Hooks, A. P., Lange, E. C., Pluer, B., and Travis, J. (2019), "Widespread biases in ecological and evolutionary studies", Bioscience, Vol. 69 No. 8, pp. 631-640.

Diakopoulos, N., and Koliska, M. (2017), "Algorithmic transparency in the news media", Digital Journalism, Vol. 5 No. 7, pp. 809-828.

Fairclough, R. and Thelwall, M. (2022), "Questionnaires mentioned in academic research 1996-2019: Rapid increase but declining citation impact", Learned Publishing, Vol. 35, pp. 241-252.

Fiez, T., Shah, N., & Ratliff, L. (2020), "A SUPER* algorithm to optimize paper bidding in peer review". In Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), (pp. 580-589).

Furl, N., Phillips, P. J., and O'Toole, A. J. (2002), "Face recognition algorithms and the other-race effect: computational mechanisms for a developmental contact hypothesis", Cognitive Science, Vol. 26 No. 6, pp. 797-815.

Gans, J. S., and Shepherd, G. B. (1994), "How are the mighty fallen: Rejected classic articles by leading economists", Journal of Economic Perspectives, Vol. 8 No. 1, pp. 165-179.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., and Yang, G. Z. (2019), "XAI—Explainable artificial intelligence", Science Robotics, Vol. 4 No. 37, pp. eaay7120.

Harzing, A. (2017), "Web of Science: How to be robbed of 10 years of citations in one week!" https://harzing.com/blog/2017/02/web-of-science-to-be-robbed-of-10-years-of-citations-in-one-week

HEFCE (2015), "The Metric Tide: Correlation analysis of REF2014 scores and metrics (Supplementary Report II to the independent Review of the Role of Metrics in Research Assessment and Management)", Higher Education Funding Council for England. https://www.ukri.org/wp-content/uploads/2021/12/RE-151221-TheMetricTideFullReport-REF2014Scores.pdf

Hicks, D., Wouters, P., Waltman, L., De Rijcke, S., and Rafols, I. (2015), "Bibliometrics: the Leiden Manifesto for research metrics", Nature, Vol. 520 No. 7548, pp. 429-431.

Hojat, M., Gonnella, J. S., and Caelleigh, A. S. (2003), "Impartial judgment by the "gatekeepers" of science: fallibility and accountability in the peer review process", Advances in Health Sciences Education, Vol. 8 No. 1, pp. 75-96.

Hook, D. W., Porter, S. J., and Herzog, C. (2018), "Dimensions: building context for search and evaluation", Frontiers in Research Metrics and Analytics, Vol. 3, pp. 23. https://doi.org/10.3389/frma.2018.00023

Hutchins, B. I., Yuan, X., Anderson, J. M., and Santangelo, G. M. (2016), "Relative citation ratio (RCR): a new metric that uses citation rates to measure influence at the article level", PLoS Biology, Vol. 14 No. 9, pp. e1002541.

Jannot, A. S., Agoritsas, T., Gayet-Ageron, A., and Perneger, T. V. (2013), "Citation bias favoring statistically significant studies was present in medical research", Journal of clinical Epidemiology, Vol. 66 No. 3, pp. 296-301.

Karhulahti, V. M., and Backe, H. J. (2021), "Transparency of peer review: a semi-structured interview study with chief editors from social sciences and humanities", Research Integrity and Peer Review, Vol. 6 No. 1, pp. 1-14.

Klavans, R., and Boyack, K. W. (2017), "Which type of citation analysis generates the most accurate taxonomy of scientific and technical knowledge?", Journal of the Association for Information Science and Technology, Vol. 68 No. 4, pp. 984-998.

Kordzadeh, N., and Ghasemaghaei, M. (2021), "Algorithmic bias: review, synthesis, and future research directions", European Journal of Information Systems, Vol. 31 No. 3, pp. 388-409.

Kumar, S., Arora, H., Ghosal, T., and Ekbal, A. (2022), “DeepASPeer: towards an aspect-level sentiment controllable framework for decision prediction from academic peer reviews”. In Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries (pp. 1-11).

Lambrecht, A., and Tucker, C. (2019), "Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads", Management Science, Vol. 65 No. 7, pp. 2966-2981.

Lancho Barrantes, B. S., Guerrero Bote, V. P., Rodríguez, Z. C., and de Moya Anegón, F. (2012), "Citation flows in the zones of influence of scientific collaborations", Journal of the American Society for Information Science and Technology, Vol. 63 No. 3, pp. 481-489.

Lee, C. J., Sugimoto, C. R., Zhang, G., and Cronin, B. (2013), "Bias in peer review", Journal of the American Society for Information Science and Technology, Vol. 64 No. 1, pp. 2-17.

Lee, N. T. (2018), "Detecting racial bias in algorithms and machine learning. Journal of Information, Communication and Ethics in Society", Vol. 16 No. 3, pp. 252-260.

Mahoney, M. J. (1977), "Publication prejudices: An experimental study of confirmatory bias in the peer review system", Cognitive Therapy and Research, Vol. 1, pp. 161-175.

Mañana-Rodríguez, J. (2015), "A critical review of SCImago journal & country rank", Research Evaluation, Vol. 24 No. 4, pp. 343-354.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2021), "A survey on bias and fairness in machine learning", ACM Computing Surveys (CSUR), Vol. 54 No. 6, pp. 1-35.

Merton, R. K. (1968), "The Matthew Effect in Science: The reward and communication systems of science are considered", Science, Vol. 159 No. 3810, pp. 56-63.

Mongeon, P., and Paul-Hus, A. (2016), "The journal coverage of Web of Science and Scopus: a comparative analysis", Scientometrics, Vol. 106 No. 1, pp. 213-228.

Moys, J. L. (2014), "Typographic layout and first impressions: testing how changes in text layout influence reader's judgments of documents", Visible Language, Vol. 48 No. 1,

pp. 881.
https://centaur.reading.ac.uk/39859/1/Typographic%20layout%20ACCEPTED.pdf
Navarro, C. L. A., Damen, J. A., Takada, T., Nijman, S. W., Dhiman, P., Ma, J., and Hooft, L. (2021), "Risk of bias in studies on prediction models developed using supervised machine learning techniques: systematic review", BMJ, Vol 2281. pp. 375.
Prates, M. O., Avelar, P. H., and Lamb, L. C. (2020), "Assessing gender bias in machine translation: a case study with google translate", Neural Computing and Applications, Vol. 32 No. 10, pp. 6363-6381.
Rowson, B., Duma, S. M., King, M. R., Efimov, I., Saterbak, A., and Chesler, N. C. (2021), "Citation diversity statement in BMES journals", Annals of Biomedical Engineering, Vol. 49 No. 3, pp. 947-949.
Ruiz-Castillo, J., and Waltman, L. (2015), "Field-normalized citation impact indicators using algorithmically constructed classification systems of science", Journal of Informetrics, Vol. 9 No. 1, pp. 102-117.
Sandström, U., and Hällsten, M. (2008), "Persistent nepotism in peer-review", Scientometrics, Vol. 74 No. 2, pp. 175-189.
Siler, K., Lee, K., and Bero, L. (2015), "Measuring the effectiveness of scientific gatekeeping", Proceedings of the National Academy of Sciences, Vol. 112 No. 2, pp. 360-365. https://www.pnas.org/doi/pdf/10.1073/pnas.1418218112
Thelwall, M. (2017), "Three practical field normalised alternative indicator formulae for research evaluation. Journal of Informetrics, 11(1), 128–151. 10.1016/j.joi.2016",12.002
Thelwall, M. (2018), "Gender bias in machine learning for sentiment analysis, Online Information Review, 42(3), 343-354. https://doi.org/10.1108/OIR-05-2017-0152
Thelwall, M. (2020), "Female citation impact superiority 1996–2018 in six out of seven English-speaking nations", Journal of the Association for Information Science and Technology, Vol. 71 No. 8, pp. 979-990.
Thelwall, M., Kousha, K., Abdoli, M., Stuart, E., Makita, M., Wilson, P., and Levitt, J. (2022), "Can REF output quality scores be assigned by AI? Experimental evidence". arXiv preprint arXiv:2212.08041.
Thelwall, M., Kousha, K., Wilson, P. Makita, M., Abdoli, M., Stuart, E., Levitt, J., Knoth, P., and Cancellieri, M. (2023), "Predicting article quality scores with machine learning: The UK Research Excellence Framework", Quantitative Science Studies, Vol. 4 No. 2, pp. 547–573.
Thelwall, M., and Maflahi, N. (2015), "Are scholarly articles disproportionately read in their own country? An analysis of Mendeley readers", Journal of the Association for Information Science and Technology, Vol. 66 No. 6, pp. 1124-1135.
Thelwall, M., and Nevill, T. (2018), "Could scientists use Altmetric", com scores to predict longer term citation counts? Journal of Informetrics, Vol. 12 No. 1, pp. 237-248.
Tincani, M., and Travers, J. (2019), "Replication research, publication bias, and applied behavior analysis", Perspectives on Behavior Science, Vol. 42 No. 1, pp. 59-75.
Traag, V. A., and Waltman, L. (2019), "Systematic analysis of agreement between metrics and peer review in the UK REF", Palgrave Communications, Vol. 5 No. 1, pp. article 29.
Tran, D., Valtchanov, A., Ganapathy, K., Feng, R., Slud, E., Goldblum, M., and Goldstein, T. (2020), "An open review of openreview: A critical analysis of the machine learning conference review process". arXiv preprint arXiv:2010.05137.

Urlings, M. J., Duyx, B., Swaen, G. M., Bouter, L. M., and Zeegers, M. P. (2021), "Citation bias and other determinants of citation in biomedical research: findings from six citation networks", Journal of Clinical Epidemiology, Vol. 132, pp. 71-78.

van Eck, N. J., and Waltman, L. (2019), "Accuracy of citation data in Web of Science and Scopus". arXiv preprint arXiv:1906.07011.

van Leeuwen, T., Moed, H., Tijssen, R., Visser, M., and Van Raan, A. (2001), "Language biases in the coverage of the Science Citation Index and its consequences for international comparisons of national research performance", Scientometrics, Vol. 51 No. 1, pp. 335-346.

Vera-Baceta, M. A., Thelwall, M., and Kousha, K. (2019), "Web of Science and Scopus language coverage", Scientometrics, Vol. 121 No. 3, pp. 1803-1813.

Vilone, G., and Longo, L. (2020), "Explainable artificial intelligence: a systematic review". arXiv preprint arXiv:2006.00093.

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., and van Raan, A. F. (2011), "Towards a new crown indicator: An empirical analysis", Scientometrics, Vol. 87 No. 3, pp. 467-481.

Wang, J., Veugelers, R., and Stephan, P. (2017), "Bias against novelty in science: A cautionary tale for users of bibliometric indicators", Research Policy, Vol. 46 No. 8, pp. 1416-1436.

Wang, Q., and Sandström, U. (2015), "Defining the role of cognitive distance in the peer review process with an explorative study of a grant scheme in infection biology", Research Evaluation, Vol. 24 No. 3, pp. 271-281.

Wilsdon, J., Allen, L., Belfiore, E., Campbell, P., Curry, S., Hill, S., and Johnson, B. (2015), "The metric tide", Report of the independent review of the role of metrics in research assessment and management". https://www.ukri.org/publications/review-of-metrics-in-research-assessment-and-management/

Wren, J. D. (2018), "Algorithmically outsourcing the detection of statistical errors and other problems", The EMBO Journal, Vol. 37 No. 12, pp. e99651.

Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., & Zhu, J. (2019), "Explainable AI: A brief survey on history, research areas, approaches and challenges". In CCF International Conference on Natural Language Processing and Chinese Computing (pp. 563-574), Berlin, Germany: Springer.

Zaharie, M. A., and Osoian, C. L. (2016), "Peer review motivation frames: A qualitative approach", European Management Journal, Vol. 34 No. 1, pp. 69-79.

Zhang, H. (2010), "CrossCheck: an effective tool for detecting plagiarism", Learned Publishing, Vol. 23 No. 1, pp. 9-14.