Received 30 August 2021
Revised 3 June 2022
Accepted 27 January 2023

A systematic review of socio-technical gender bias in AI algorithms

Paula Hall and Debbie Ellis
School of Management, IT and Governance, University of KwaZulu-Natal - Pietermaritzburg Campus, Pietermaritzburg, South Africa

Abstract

Purpose – Gender bias in artificial intelligence (AI) should be solved as a priority before AI algorithms become ubiquitous, perpetuating and accentuating the bias. While the problem has been identified as an established research and policy agenda, a cohesive review of existing research specifically addressing gender bias from a socio-technical viewpoint is lacking. Thus, the purpose of this study is to determine the social causes and consequences of, and proposed solutions to, gender bias in AI algorithms.

Design/methodology/approach – A comprehensive systematic review followed established protocols to ensure accurate and verifiable identification of suitable articles. The process revealed 177 articles in the socio-technical framework, with 64 articles selected for in-depth analysis.

Findings – Most previous research has focused on technical rather than social causes, consequences and solutions to AI bias. From a social perspective, gender bias in AI algorithms can be attributed equally to algorithmic design and training datasets. Social consequences are wide-ranging, with amplification of existing bias the most common at 28%. Social solutions were concentrated on algorithmic design, specifically improving diversity in AI development teams (30%), increasing awareness (23%), human-in-the-loop (23%) and integrating ethics into the design process (21%).

Originality/value – This systematic review is the first of its kind to focus on gender bias in AI algorithms from a social perspective within a socio-technical framework. Identification of key causes and consequences of bias and the breakdown of potential solutions provides direction for future research and policy within the growing field of AI ethics.

Peer review – The peer review history for this article is available at https://publons.com/publon/10.1108/OIR-08-2021-0452

Keywords AI gender Bias, Algorithmic bias, Gender bias, Socio-technical framework, AI ethics

Paper type Research paper

Introduction

Artificial intelligence (AI) is rapidly and fundamentally changing many business models, industries, and the nature of work itself. At the core of this change are predictive algorithms, which are developed predominantly by men, and trained on data sets with inherent gender biases (Bourton *et al.*, 2018; Foulds *et al.*, 2020). Some early AI applications had to be withdrawn due to gender bias. Examples of this are Amazon and Linked-In's AI-based resume services that only sent information technology (IT) job openings to male job seekers (Varghese, 2019; West *et al.*, 2019). Research has supported these anecdotal biases, for example, a recent study entitled "Computer algorithms prefer headless women" showed how advertising algorithms are demonstrating sexist tendencies when it comes to images used in online targeted marketing (Cecere *et al.*, 2018). One of the manifestations of underrepresentation of women in training data sets and in AI development teams is a bias against women in AI solutions, perpetuating gender imbalances, often in opposition to international legal and ethical management practices (Bourton *et al.*, 2018; Baumann and Rumberger, 2018; Yapo and Weiss, 2018).

Online Information Review
Vol. 47 No. 7, 2023
pp. 1264-1279

1468-4527
DOI 10.1108/OIR-08-2021-0452

As evidence of the consequences of AI bias mounts, there have been calls to increase algorithmic transparency (Thelwall, 2018) and hold the producers of AI applications accountable (Raji and Buolamwini, 2019) as part of a growing AI ethics movement. Algorithmic bias has caused real harm to marginalised elements of society, including women and non-binary minorities (Saka, 2020, Leavy *et al.*, 2020). There are technical and societal causes for this, and solutions should be found within a socio-technical framework (Draude *et al.*, 2019; Selbst *et al.*, 2019). There is no comprehensive "stock take" of the extent and nature of the problem which details the causes, consequences, and proposed solutions, such as in the form of a systematic review. Thus, the purpose of this study is to conduct a systematic review of the literature to investigate gender bias in AI algorithms. Specifically the objectives are to

(1) broadly categorise social causes of gender bias in AI algorithms.

(2) outline the social consequences of algorithmic gender bias.

(3) explore proposed social solutions to AI gender bias.

This paper proceeds as follows: first a literature review is presented, including the socio-technical framework which provides the theoretical foundation for this study, and ending with an overview of related systematic reviews. Then the PRISMA methodology and data collection for the systematic review are discussed. Thereafter, the data analysis and results are presented. Finally, the discussion and recommendations conclude the paper.

Literature review

Algorithms are key to a subset of AI called machine learning, used in industries as diverse as education, finance, marketing, transport and security (Akerkar, 2019; Bucher, 2018). Algorithms discover patterns in large datasets from which they learn rules for automated predictions and decision-making (Bucher, 2018). Since algorithms are designed to discriminate, in that they sort, classify and make inferences about data (Veale and Binns, 2017), biases contained in historical data and machine learning models become part of the solution (World Economic Forum, 2018). Additionally, how the algorithms come to the answers they do is often opaque, not explainable and proprietary (Burrell, 2016). These problems will become more pressing with increasingly complex patterns in data, sophisticated multi-level neural networks and more autonomy in algorithms (Osoba and Welser, 2017).

Algorithmic bias refers to unfair discrimination against gender, race and other groups (Springer *et al.*, 2018). Socially acceptable inferences can be programmed in to ensure fairness, but this will result in conflict with the accuracy requirements of algorithm outputs (Osoba and Welser, 2017). Further problems arise from models used in algorithms being able to implement all the measures of fairness simultaneously (Veale and Binns, 2017). Subjective human involvement comes into the selection of features and models used in the algorithms (Veale and Binns, 2017) and how they include identifiable factors, such as race and gender.

A socio-technical framework

Several technical approaches to identifying and reducing algorithmic bias have been proposed, including using statistical methods, audits of algorithms, automated reasoning and transparency and explainability requirements (Osoba and Welser, 2017; Saka, 2020; Veale and Binns, 2017; Burrell, 2016). Non-technical approaches recommended by Burrell (2016) and Osoba and Welser (2017) include user literacy and education, diversity in algorithm developers and algorithm regulation. Veale and Binns (2017) and Selbst *et al.* (2019) argue that

a socio-technical approach is required, given the complexity and context-specific nature of the issue. A socio-technical analysis was also presented by Monteiro *et al.* (2021) to minimise potential algorithmic bias.

Draude *et al.* (2019) developed a multi-disciplinary, socio-technical framework used in this review. This framework presents a socio-technical gender perspective, consisting of data bias, technical model bias and emergent bias. Data bias is shaped by the society generating the data, which is fed into the algorithms, while technical bias arises from the models and features of the algorithm itself; and emergent bias stems from the way the algorithm is used (Draude *et al.*, 2019). Table 1 below reflects these key concepts in the socio-technical framework adopted in this review.

Previous systematic reviews on the topic of AI bias

A systematic review on perceptions of fairness in AI was conducted by Baleis *et al.* (2019) which included some elements of AI bias but was focused on people's cognitive and emotional responses to, and perceptions of the issue of fairness in AI. Baleis *et al.* (2019) found people's perceptions of algorithmic fairness differed based on context and recommend further interdisciplinary research. The review focused on perceptions of bias rather than causes, consequences and solutions. Khalil *et al.* (2020) conducted a systematic review into bias in facial recognition systems, which is a specific application of AI. The review considered all forms of bias, including racial, gender, age, culture and ethnicity, and found the main cause was the lack of variability in training data sets (Khalil *et al.*, 2020). Köchling and Wehner (2020) also conducted a review in a specific application area – that of algorithmic bias in human resources, finding bias resulting from use of AI in recruitment and human resource development.

The related topic of algorithmic accountability was dealt with in a systematic review by Wieringa (2020) who used accountability theory to review the elements of accountability in algorithms from a socio-technical viewpoint. Wieringa (2020) was able to define and integrate accountability theory into algorithmic accountability from a multi-disciplinary perspective. The systematic review in this paper also takes a multi-disciplinary approach; however, it is focused on algorithmic gender bias rather than accountability. Another review, by Favaretto *et al.* (2019), focused on discriminatory risks of data mining and associated biases and found human bias is a barrier to fair AI systems. The study concluded that discrimination in data mining is significant, yet often underestimated, and that more research is needed into discriminatory practices in big data (Favaretto *et al.*, 2019).

None of the above reviews focused specifically on gender bias social causes, consequences and solutions. Similar to Wieringa (2020), this review also adopts a socio-technical framework.

Table 1.
Key concepts in the socio-technical framework

Label	Description
Datasets	Ensuring that data sets do not include gender unless needed, and that proxy variables for gender are also excluded, or that there is awareness and oversight of data collection and classification to prevent gender discrimination (Draude *et al.*, 2019). To counter data bias
Algorithm design	Increasing gender diversity in the entire design lifecycle of an Al solution (Draude *et al.*, 2019). To counter technical model bias
Context	Considering the complete context in which the algorithm is situated, from the stakeholders, to the data set to the socio-technical systems (Draude *et al.*, 2019). To counter emergent bias
Due process	Address fairness, accountability, and transparency factors from a development process and policy viewpoint (Draude *et al.*, 2019). To counter gender bias in general
Other	Other causes, for results that cannot be classified according to the other labels

Research methodology

A systematic review is an exacting methodology for finding, analysing and reporting literature in a specific precisely defined area producing a replicable and transparent synthesis of the current state of knowledge (Denyer and Tranfield, 2009). By following a rigorous scientific approach, relevant literature can be organised and evidenced-based decisions made about knowledge gaps (Petticrew and Roberts, 2008) in the field of gender bias in AI algorithms. The key principles of systematic reviews include a lack of bias by using a well-defined protocol to ensure validity, rigour and replicability (Shamseer *et al.*, 2015). In this way the results can be used with confidence to inform practice, policy and future research direction (Denyer and Tranfield, 2009). By specifically targeting gender bias, a homogenous set of results can be expected from this review, allowing for aggregation and synthesis of results. Petticrew and Roberts (2008) argue that systematic reviews provide evidence-based analysis and synthesis of previous studies, creating an overall picture of the topic. In the area of software engineering, systematic reviews offer a clear picture of the latest, state of the art in research (Kitchenham and Brereton, 2013) which is pertinent to this topic on gender bias in AI algorithms. Rather than a primary study that would only add more detail to a narrow area of the topic, a systematic review provides a precursor to further research (Petticrew and Roberts, 2008), in this case on AI gender bias, so that the problem can move forward to finding potential solutions. Arguments that systematic reviews are not useful in an immature field such as gender bias in AI algorithms are countered by Petticrew and Roberts (2008) who suggest that such reviews are suitable to identify gaps, research directions and potential early interventions.

Research question

The research question for this review has been constructed using the Population, Intervention, Comparison, Outcomes and Context (PICOC) framework for systematic review research questions (Petticrew and Roberts, 2008) which provides a formal and specific method for deriving a research question. The focus of the question is gender bias in AI algorithms. Similar to the process used by Baleis *et al.* (2019), the application of the PICOC framework for this study is: the Population being women, the Intervention/Indicator is gender bias in AI algorithms, with the Outcome being the effect of gender bias in a global context. Comparison was not applicable. From this the following research question was derived:

What are the social causes and consequences of, and solutions to gender bias in AI algorithms? The following sub-questions arise:

(1) What is the breakdown of technical versus social approaches to algorithmic gender bias?

(2) What are the social causes of gender bias in AI algorithms?

(3) What are the social consequences of algorithmic gender bias?

(4) What social solutions are proposed for AI gender bias?

Data collection

The data collection procedure followed a strict protocol to ensure explicit and transparent description and justification of the procedure followed. The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) protocol was selected since it is the preferred method for ensuring a rigorous review, and is used by other systematic reviews on the topic of algorithms (e.g. Wieringa, 2020; Favaretto *et al.*, 2019).

A scoping review using Google Scholar determined the size and scope of the review, as recommend by Tranfield *et al.* (2003), due to the cross-disciplinary nature of the topic.

This exercise informed the search strategy for the review, including search terms, type of literature and study period (Linnenluecke *et al.*, 2019). The results also determined that this review was not duplicating an existing review on algorithmic gender bias. In their systematic review on fairness and big data, Favaretto *et al.* (2019) cite the lack of grey literature in their review as a limitation, and recommend that future studies include some forms of this. For this reason, and based on the scoping review, grey literature in the form of conference proceedings and reports were included in this review.

Kitchenham and Brereton (2013) and Adams *et al.* (2017) recommend the following databases for inclusion in business and software engineering systematic reviews: IEEE, ACM, SCOPUS, Web of Science, EBSCOHost and PsychInfo. Selection from this list for this review was guided by reviewing the databases used by systematic reviews in the general field of AI fairness and bias (Favaretto *et al.*, 2019; Khalil *et al.*, 2020; Wieringa, 2020) resulting in the final selection of IEEE Xplore, ACM Digital Library, SCOPUS and Web of Science for this review. The search terms were sourced from the focus area of AI algorithms and the factor of interest being gender bias, a similar process followed in the review by Khalil *et al.* (2020). Synonyms for keywords were derived from key studies on the issue (Wieringa, 2020). Boolean operators were used to combine the keywords from which the following search string was generated:

> ("gender bias" OR sexist OR "gender discrimination" OR "gender diversity" OR "gender inequality") AND (Algorithm OR Algorithmic OR "machine learning")

Inclusion and exclusion criteria

According to Shamseer *et al.* (2015), the PRISMA protocol specifies two types of eligibility. The first stems from characteristics of the study itself, in this case, eligible studies must be focus on gender bias, the gender bias must arise from machine learning algorithms and the study must include cause or consequences of the bias or offer potential solutions. The second stems from the characteristics of the report, including time frame, language, and publication types. Large scale deployment of machine learning algorithms is a relatively recent development (OECD, 2019) and the consequences of bias have only become a pressing issue over the last decade (World Economic Forum, 2018). Therefore, similar to other systematic reviews on the issue of AI bias (Favaretto *et al.*, 2019; Baleis *et al.*, 2019), this review focused on papers from the last 10 years. Languages other than English were excluded, supporting replicability (Wieringa, 2020), expediency and the multi-disciplinary nature of the topic. There were no restrictions based on the discipline. Publication types were restricted to journals, conference proceedings and think tank reports. There were no exclusions based on methodology, allowing empirical, theoretical and conceptual papers to be included.

Data extraction and analysis

Data extraction was conducted through the use of Distiller SR, software designed to facilitate systematic reviews (Shamseer *et al.*, 2015). The initial search produced 3,401 unique papers, with 2,752 excluded after initial screening of the title and abstract. A further 471 papers were excluded upon further review of the full paper, mainly for not including gender bias specifically. The resulting 177 papers covered both technical and social causes, consequences and solutions. Only social aspects were considered for this review, resulting in a further exclusion of 112 technical only papers, leaving 57 papers. Through following references, based on seminal papers and the Pearl or snowball method, an additional 8 papers were added, resulting in a final total of 64 papers. The results of this process can be seen in Figure 1, the PRISMA flowchart.

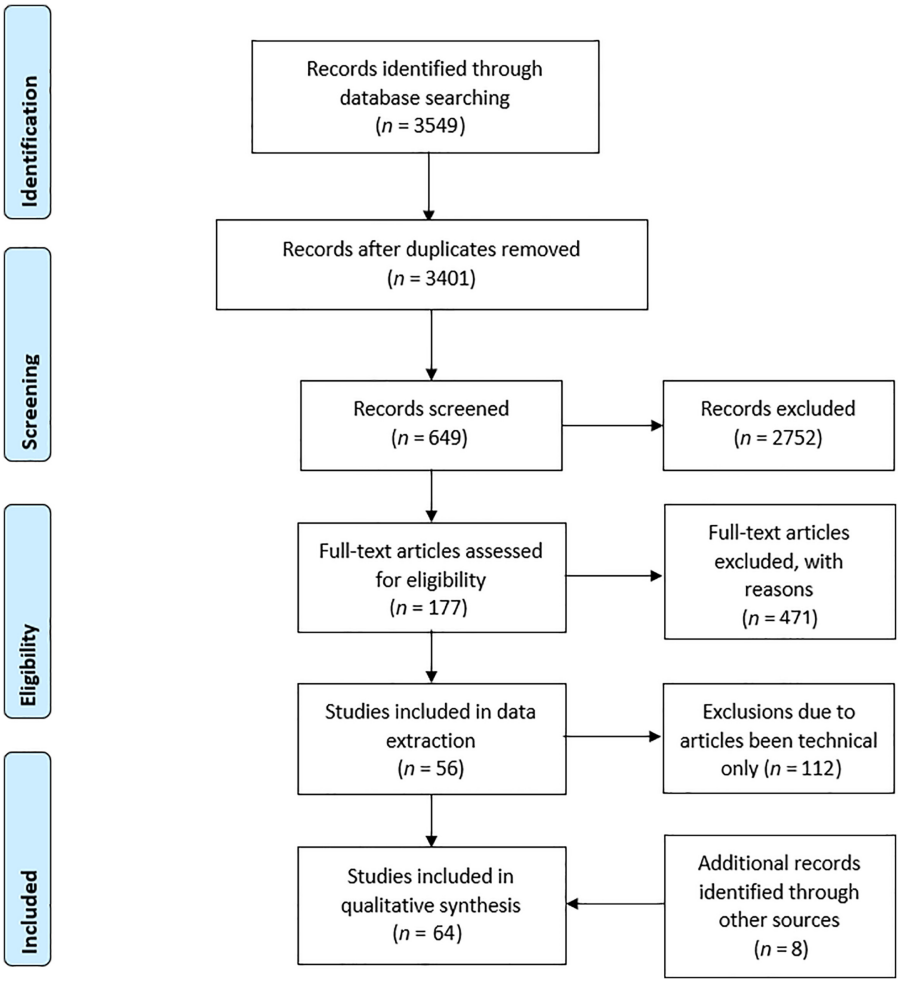

Figure 1.
The PRISMA flowchart

Data analysis followed a deductive approach, based on the literature review and research questions, which yielded themes suitable for a thematic analysis (Fereday and Muir-Cochrane, 2006). The three themes followed the code book structure presented by Fereday and Muir-Cochrane (2006) were social causes, consequences and solutions related to gender bias in AI algorithms. Direct-quote evidence of the themes were captured and categorised in the Distiller SR software and then imported into Nvivo for rich, in-depth analysis of the themes.

Quality control

Articles included for review were screened for quality of reporting and methodology (Kitchenham and Brereton, 2013). Distiller SR software was used to ensure rigour in the process of selection, inclusion and exclusion of quality articles that met the research criteria. As an additional layer of quality control, the DistillerSR AI toolkit, was used to assist in

validity and optimisation of screening systematic reviews (Hamel *et al.*, 2020). This AI screening of the manual reviews suggested 28 exclusion discrepancies, which were then manually checked. The second reviewer reviewed the methodology and evaluation process and tested a portion of randomly selected articles from each level of review as an additional level of reliability.

Results

The initial screening included social and technical papers. There were 112 papers excluded for being technical only, with 21 of the final 64 included articles falling under both social and technical categories. Figure 2 shows the breakdown of papers along the socio-technical dichotomy:

Of the included 64 papers, 31 were coded as including causes, 33 as consequences and 43 as solutions. Twenty nine of the articles also referenced racial bias, among other types of bias. Despite searching for articles from 2010, Figure 3 shows that all of the included studies ranged from 2015 onwards, with 30 from 2020 alone, and 75% from 2019 to 2020, signifying the recency of the topic. This also explains why over 50% of the articles were conference papers.

The articles were varied in domains, with 72% stemming from the field of computer science, 25% from the social sciences and the remaining 3% from the health sciences.

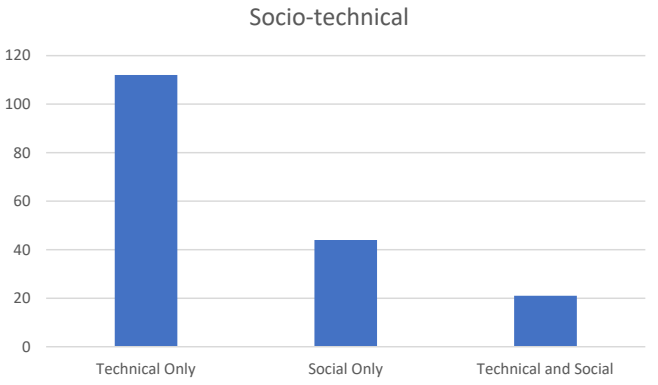

Figure 2.
Social and technical focus of included papers

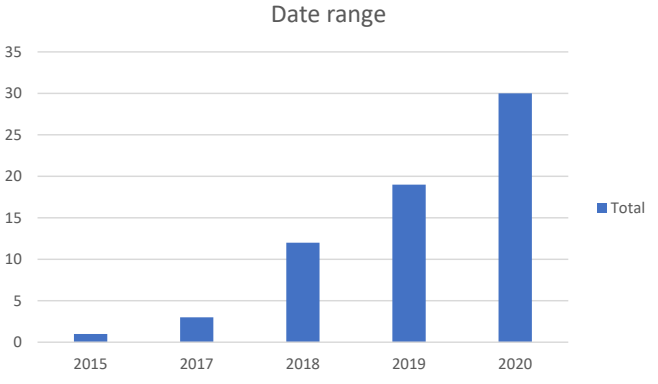

Figure 3.
Date range of systematic review articles

Specific domains reflect these general fields, with around 50% in AI ethics and AI. The remaining subject areas are shown in Figure 4 below.

Social causes of algorithmic gender bias

For the social causes of gender bias, the articles were equally attributed to the design of AI algorithms and to the datasets they are trained on. Figure 5 shows that within the algorithmic design category, 43% were related to the lack of diversity in development teams, while 36% specified lack of awareness of bias in the algorithmic design process.

Quotes from articles to illustrate examples of findings are presented in italics: For example Avellan *et al.* (2020) state "*homogeneous teams will share the same blind spots or cognitive biases that will transfer to their design of the technology, creating unbalanced and unfair outcomes*." Bias is introduced at many points, including the design process (Singh *et al.*, 2020) with the many roles in AI development teams demonstrating a lack of diversity (Dillon and Collett, 2019). Ntoutsi *et al.* (2020) suggest that "*representation-related biases creep into development processes because the development teams are not aware of the importance of distinguishing between certain categories*" and that lack of diversity in development teams is one of the causes.

For the datasets cause, the majority of articles (85%) referenced the lack of diversity in datasets. This is outlined by Leavy *et al.* (2020) "*The source of this kind of bias often lies in the way societal inequalities and latent discriminatory attitudes are captured in the data from which*

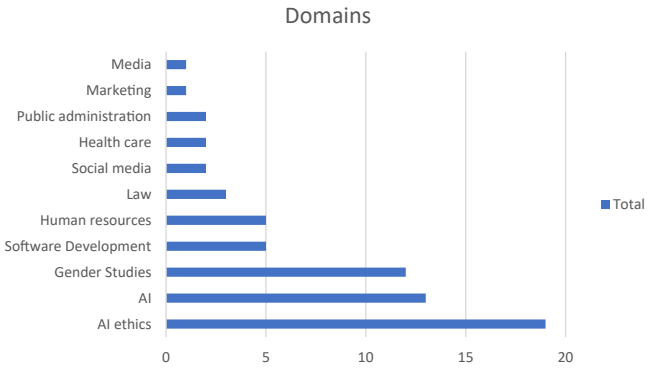

Figure 4.
Domains represented by systematic review articles

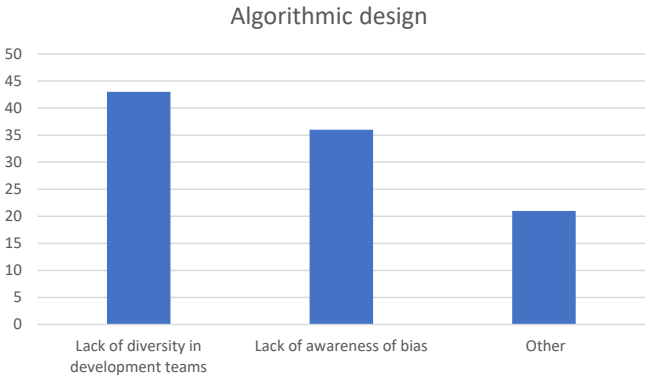

Figure 5.
Sub-themes for theme "algorithmic design" for social causes

algorithms learn". Wellner (2020) agrees that gender bias arises from the training data, while Saka (2020) explains that algorithms trained on big data containing gender stereotypes and underrepresented gender minorities will perpetuate those biases in the future. The "Other" theme included gender bias arising from incorrect classification of data, often at the intersection of race and gender (Buolamwini and Gebru, 2018).

Social consequences of algorithmic gender bias

Social consequences of gender bias in AI algorithms were wide ranging in the articles as depicted in Figure 6.

The most commonly cited consequence was the amplification of gender bias, often creating a feedback loop: "*Due to the feedback loop mechanism, the gender-biased results are fed back to the system, thereby deepening the biases*" (Wellner, 2020). The next most commonly coded consequence, at 25%, was that search results were biased, which included text, translation and image search results, most commonly from Google search. Leavy *et al.* (2020) suggest that evidence of gender bias in recommender and search systems will have a profound impact on society, given the pervasiveness of these tools in everyday use. Otterbacher *et al.* (2017) found a similar issue with Bing, and how its image search algorithm tends to produce results that reinforce gender bias. Search text and translation results often illustrated gender bias in profession or occupation: "*We then show that Google Translate exhibits a strong tendency towards male defaults, in particular for fields typically associated to unbalanced gender distribution or stereotypes such as STEM (Science, Technology, Engineering and Mathematics) jobs*" (Prates *et al.*, 2020). Gender bias had consequences in many fields, those that came up most frequently were criminal justice (15%), hiring (10%) and advertising (7%). Criminal justice tends to manifest intersectionally, with black women the most adversely affected (Fernández-Martínez and Fernández, 2020; Hamilton, 2019). In hiring, Leicht-Deobald *et al.* (2019) explain that recruitment algorithms may be "actively reifying the original gender bias" while Google's advertising algorithm tends to bypass non-binary gender choices when personalising adverts (Shekhawat *et al.*, 2019). In the "Other" category, bias was found in the areas of facial recognition, education, automated personal assistants and surveillance (Whittaker *et al.*, 2018).

Social solutions to algorithmic gender bias

The final category considered papers proposing social solutions, with Figure 7 showing the emergent themes: 57% coded as improving algorithmic design and 33% as improving

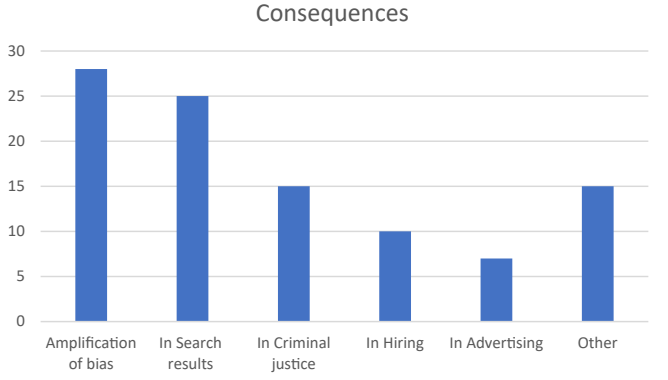

Figure 6.
Social consequences

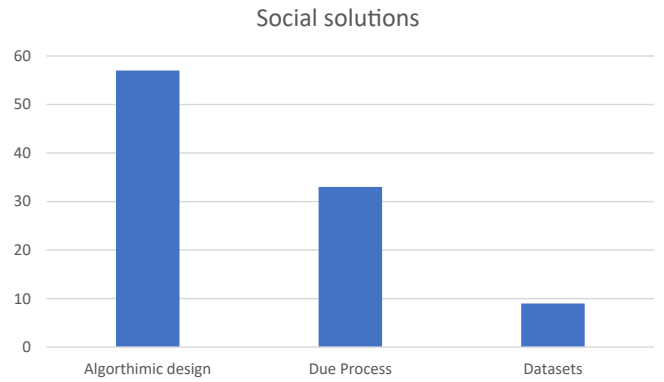

Figure 7.
Overall results for social solutions

due process. The remainder fell under the improving diversity in datasets category, with the solution emerging as ensuring the data is representative (Avellan *et al.*, 2020) and that gender bias is removed in cleaning up datasets (Parsheera, 2018).

Within the algorithmic design theme, 4 main sub-themes emerged. These are depicted in Figure 8 below.

Improving diversity in development teams was the most commonly coded solution (30%). For example: "*increasing gender inclusion in the development of AI technologies will introduce important and diverse perspectives, reduce the influence of cognitive biases in the design, training, and oversight of learning algorithms, and, thereby, mitigate bias-related risk management concern*" (Johnson, 2019). Saka (2020) contends that recent studies on gendered algorithmic bias show that increasing the number of diverse designers and developers of algorithms can mitigate discriminatory outcomes.

Increasing awareness of bias within the algorithmic design and development process (23%) is demonstrated by Wellner and Rothman (2020): "*Users and developers should be aware of the possibility of gender and racial biases, and try to avoid them, bypass them, or exterminate them altogether*".

Human-in-the-loop refers to the importance of humans in automated systems, especially in regulating AI (Rahwan, 2018) a term gaining prominence in socio-technical solutions. The human-in-the-loop solution coded at 23%, illustrated by Gilbert and Mintz (2019) who urge

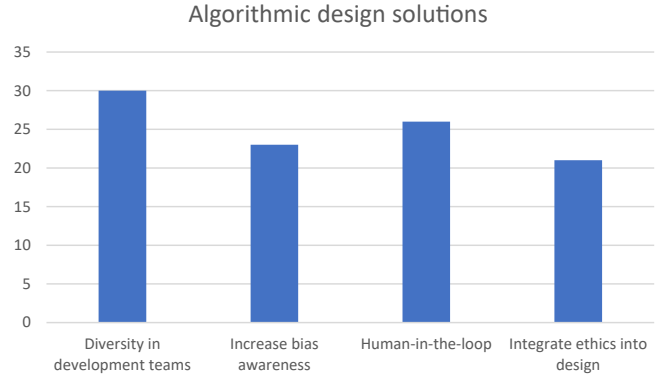

Figure 8.
Social solutions related to algorithmic design

"*engineers to interpret themselves as part of this context, which includes the wider machine learning community as well as potentially-vulnerable populations of protected social categories*".

A final solution that emerged under the algorithm design sub-theme was to integrate ethics into the algorithm design process, at 21%, by creating and implementing a gender-inclusive code of AI ethics (Badaloni and Lisl, 2020).

The sub-theme of due process made up 33% of the social solutions to gender bias, with the codes emerging shown in Figure 9.

In the due process theme, 36% were aimed at improving fairness, accountability and transparency. Fairness is illustrated by Weyerer and Langer (2019) "*the development of strict and fair decision rules of AI applications that limit any conclusion based on non-relevant personal characteristics such as race, gender, etc. is vital to prevent AI-based discrimination*". Regarding accountability, Raji *et al.* (2020) is an example of a study where they "*contribute to closing the accountability gap in the development and deployment of large-scale artificial intelligence systems*". Transparency is a key recommendation by West *et al.* (2019): "*Remedying bias in AI systems is almost impossible when these systems are opaque. Transparency is essential, and begins with tracking and publicizing where AI systems are used, and for what purpose*". Monteiro *et al.* (2021) agree, suggesting that transparency by design is required in preventing bias. Auditing as a due process solution to algorithmic bias (20%) is supported by Raji *et al.* (2020) who created a "*framework for algorithmic auditing that supports artificial intelligence system development end-to-end, to be applied throughout the internal organization development lifecycle*". Legal regulations to address algorithmic bias made up 28% of the coding. For example, Busuioc (2020) suggested "*Regulatory efforts are thus vitally needed to ensure that AI tools are brought to bear in a thoughtful and effective manner*".

Discussion

The socio-technical framework illustrates the importance of viewing issues such as AI bias from both a social and technical viewpoint. However the results of this review show that most previous research has focused on the technical aspects only, a narrow approach that has drawn criticism (Whittaker *et al.*, 2018). While success of AI systems requires a combined socio-technical perspective, the human or social side should be the driver (Singh *et al.*, 2020). Given the multidisciplinary nature of the studies in this review, a social view is appropriate given how many sectors and actors are affected (Raisch and Krakowski, 2021).

Within the social context, the results demonstrate that gender bias can be introduced at multiple points, from the datasets, algorithm design and use of AI (Singh *et al.*, 2020). This is

Figure 9.
Social solutions related to due process

corroborated by the results of the causes theme. Causes related to training data bias and lack of gender diversity in programming teams are corroborated empirically in a study by Cowgill *et al.* (2020).

Although all consequences of AI gender bias emerging from this review are serious, of concern is the most cited consequence: amplification via the feedback loop. "*Algorithmic bias has the capacity to amplify and perpetuate societal bias, and presents profound ethical implications for society*" (Leavy *et al.*, 2020). The consequences emerging from the studies included in this review are far-reaching, multidisciplinary, and touch multiple industries and fields. This in turn creates further "*discrimination and invisibilisation of women*" (Gutierrez, 2021).

Social solutions link back to the social causes, for example diversity in AI development teams emerged as a leading cause and solution. This is corroborated by several recent studies, with Dillon and Collett (2019) contending that "Diversification of the AI workforce will be vital in order to design and implement technology which is equitable". UNESCO (2020) recommend inclusiveness, and parity for women in AI development teams. As expected, another social solution around algorithmic design emerged as ensuring human oversight, classified as humans-in-loop. For instance, Bowen *et al.* (2020) recommend more cross-disciplinary "designers in the loop". Further, the "right" humans are needed for intersectional representation (UNESCO, 2020) and socially responsible algorithms (Cheng *et al.*, 2021).

The solutions start with acknowledgement and recognition that simply improving awareness of the problem is a good start. Leicht-Deobald *et al.* (2019) suggest that ethical awareness is a key aspect of overcoming AI bias challenges. Having human oversight extends to the due process category. There have been many studies dedicated to fairness, accountability, transparency and auditing of AI algorithms to mitigate bias (UNESCO, 2020; Jobin *et al.*, 2019; Thelwall, 2018). After reviewing 24 studies, Khalil *et al.* (2020) determined that further research was needed into auditing of algorithms and benchmarking databases. Good governance, public and private sector policy and regulatory frameworks are being actively pursued in research and practice (West *et al.*, 2019; Dillon and Collett, 2019) as part of the growing field of AI ethics.

Interestingly, many of the emerging solutions from this review have been the subject of calls for further research (Dillon and Collett, 2019; UNESCO, 2020) and action, indicating the need to pursue these social solutions (Cowgill *et al.*, 2020).

Conclusion and recommendations

A more nuanced approach is needed when considering the issue of algorithmic gender bias, as the results of this review show that bias exists intersectionally, particularly for black females (Buolamwini and Gebru, 2018; Wang *et al.*, 2022). Further, gender is not a binary issue, and further research and solutions are needed to ensure truly inclusive algorithmic decision making, accounting for all sub-groups (UNESCO, 2020). Future research should therefore consider gender bias intersectionally and include gender minorities.

The issue of AI bias should be considered from a social, interdisciplinary perspective (Cheng *et al.*, 2021), as recommended by Whittaker *et al.* (2018) "*expand the disciplinary makeup of those engaged in AI design, development, and critique, beyond purely technical expertise*". Treating algorithmic bias as a technical issue and focusing on debiasing training data is too simplistic and can still create discriminatory outcomes (Marda, 2021). Further, a techno-centric view fails to consider the important role of humans in the various stages of the AI lifecycle and the consequences for society (Marda, 2021). More women and gender minorities should be included in AI management and development teams to increase awareness, improve design and advance diversity and inclusion (Smith and Rustagi, 2021). Human oversight of algorithms in use, especially in critical decisions is recommended

(Köchling and Wehner, 2020). Increasing diversity in development teams emerged as a leading solution, therefore ways to advance women's careers in AI should be prioritised for further research (Leavy, 2018; Roopaei *et al.*, 2021). This is particularly important given the structural and persistent gender inequality in AI (Young *et al.*, 2021). Ensuring stereotypes of the past are not perpetuated in the solutions (World Economic Forum, 2018; OECD, 2019) will require effort from business and society, in the form of company and public policy with a focus on human-centred, ethical AI (OECD, 2019).

References

Adams, R.J., Smart, P. and Huff, A.S. (2017), "Shades of grey: guidelines for working with the grey literature in systematic reviews for management and organizational studies", *International Journal of Management Reviews*, Vol. 19, pp. 432-454.

Akerkar, R. (2019), *Artificial Intelligence for Business*, Springer, Switzerland.

Avellan, T., Sharma, S. and Turunen, M. (2020), "AI for all: defining the what, why, and how of inclusive AI", *ACM International Conference Proceeding Series*, pp. 142-144.

Badaloni, S. and Lisl, F.A. (2020), "Towards a gendered innovation in AI", *CEUR Workshop Proceedings*, 2776, pp. 12-18.

Baleis, J., Keller, B., Starke, C. and Marcinkowski, F. (2019), "Cognitive and emotional response to fairness in AI - a systematic review", available at: https://www.sozwiss.hhu.de/fileadmin/redaktion/Fakultaeten/Philosophische_Fakultaet/Sozialwissenschaften/Kommunikations-_und_Medienwissenschaft_I/Dateien/Baleis_et_al.__2019__Literatur_Review.pdf (accessed 20 January 2021).

Baumann, E. and Rumberger, L.J. (2018), "State of the art in fair ML: from moral philosophy and legislation to fair classifiers", *arXiv Preprint*, arXiv:1811.09539.

Bourton, S., Lavoie, J. and Vogel, T. (2018), *Will Artificial Intelligence Make You a Better Leader?*, McKinsey & Company, Vol. 2, pp. 72-75.

Bowen, Y., Ye, Y., Loren, T., Zhiwei Steven, W., Jodi, F. and Haiyi, Z. (2020), "Keeping designers in the loop: communicating inherent algorithmic trade-offs across multiple objectives", *Proceedings of the 2020 ACM Designing Interactive Systems Conference*, pp. 1245-1257.

Bucher, T. (2018), *If… Then: Algorithmic Power and Politics*, Oxford University Press, New York.

Buolamwini, J. and Gebru, T. (2018), "Gender shades: intersectional accuracy disparities in commercial gender classification", *Conference on fairness, accountability and transparency*, New York, Proceedings of Machine Learning Research.

Burrell, J. (2016), "How the machine 'thinks': understanding opacity in machine learning algorithms", *Big Data and Society*, Vol. 3, pp. 1-12.

Busuioc, M. (2020), "Accountable artificial intelligence: holding algorithms to account", *Public Administration Review*, Vol. 81 No. 5, pp. 825-836.

Cecere, G., Jean, C., Manant, M. and Tucker, C. (2018), "Computer algorithms prefer headless women", *MIT CODE: Conference on Digital Experimentation*, Boston.

Cheng, L., Varshney, K.R. and Liu, H. (2021), "Socially responsible AI algorithms: issues, purposes, and challenges", *arXiv E-Prints*, arXiv: 2101.02032.

Cowgill, B., Dell'acqua, F., Deng, S., Hsu, D., Verma, N. and Chaintreau, A. (2020), "Biased programmers? Or biased data? A field experiment in operationalizing ai ethics", *Proceedings of the 21st ACM Conference on Economics and Computation*, pp. 679-681.

Denyer, D. and Tranfield, D. (2009), "Producing a systematic review", in Buchanan, D. and Bryman, A. (Eds), *The Sage Handbook of Organizational Research Methods*, Sage Publications, London.

Dillon, S. and Collett, C. (2019), *AI and Gender: Four Proposals for Future Research*, The Leverhulme Centre for the Future of Intelligence, Cambridge.

Draude, C., Klumbyte, G., Lücking, P. and Treusch, P. (2019), "Situated algorithms: a sociotechnical systemic approach to bias", *Online Information Review*, Vol. 44, pp. 325-342.

Favaretto, M., De Clercq, E. and Elger, B.S. (2019), "Big Data and discrimination: perils, promises and solutions. A systematic review", *Journal of Big Data*, Vol. 6, p. 12.

Fereday, J. and Muir-Cochrane, E. (2006), "Demonstrating rigor using thematic analysis: a hybrid approach of inductive and deductive coding and theme development", *International Journal of Qualitative Methods*, Vol. 5, pp. 80-92.

Fernández-Martínez, C. and Fernández, A. (2020), "AI and recruiting software: ethical and legal implications", *Paladyn*, Vol. 11, pp. 199-216.

Foulds, J.R., Islam, R., Keya, K.N. and Pan, S. (2020), "An intersectional definition of fairness", in *2020 IEEE 36th International Conference on Data Engineering (ICDE)*, IEEE Computer Society, pp. 1918-1921.

Gilbert, T. and Mintz, Y. (2019), "Epistemic therapy for bias in automated decision-making", *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 61-67.

Gutierrez, M. (2021), "Algorithmic gender bias and audiovisual data: a research agenda", *International Journal of Communication*, Vol. 15, pp. 439-461.

Hamel, C., Kelly, S.E., Thavorn, K., Rice, D.B., Wells, G.A. and Hutton, B. (2020), "An evaluation of DistillerSR's machine learning-based prioritization tool for title/abstract screening – impact on reviewer-relevant outcomes", *BMC Medical Research Methodology*, Vol. 20 No. 1, doi: 10.1186/s12874-020-01129-1.

Hamilton, M. (2019), "The sexist algorithm", *Behavioral Sciences and the Law*, Vol. 37, pp. 145-157.

Jobin, A., Ienca, M. and Vayena, E. (2019), "The global landscape of AI ethics guidelines", *Nature Machine Intelligence*, Vol. 1, pp. 389-399.

Johnson, K.N. (2019), "Automating the risk of bias", *George Washington Law Review*, Vol. 87, pp. 1214-1271.

Khalil, A., Ahmed, S.G., Khattak, A.M. and Al-Qirim, N. (2020), "Investigating bias in facial analysis systems: a systematic review", *IEEE Access*, Vol. 8, pp. 130751-130761.

Kitchenham, B. and Brereton, P. (2013), "A systematic review of systematic review process research in software engineering", *Information and Software Technology*, Vol. 55, pp. 2049-2075.

Köchling, A. and Wehner, M.C. (2020), "Discriminated by an algorithm: a systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development", *Business Research*, Vol. 13, pp. 795-848.

Leavy, S. (2018), "Gender bias in artificial intelligence: the need for diversity and gender theory in machine learning", *Proceedings of the 1st International Workshop on Gender Equality in Software Engineering*, Sweden, ACM, pp. 14-16.

Leavy, S., Meaney, G., Wade, K. and Greene, D. (2020), "Mitigating gender bias in machine learning data sets", *Communications in Computer and Information Science. CCIS*, Vol. 1245, pp. 12-26.

Leicht-Deobald, U., Busch, T., Schank, C., Weibel, A., Schafheitle, S., Wildhaber, I. and Kasper, G. (2019), "The challenges of algorithm-based HR decision-making for personal integrity", *Journal of Business Ethics*, Vol. 160, pp. 377-392.

Linnenluecke, M., Marrone, M. and Singh, A. (2019), "Conducting systematic literature reviews and bibliometric analyses", *Australian Journal of Management*, Vol. 45, pp. 175-194.

Marda, V. (2021), "AI bias is not just a data problem", *Forbes. India*, Forbes Media.

Monteiro, L., Zaman, B., Caregnato, S.E., Geerts, D., Grassi-Filho, V. and Htun, N.-N. (2021), "Trespassing the gates of research: identifying algorithmic mechanisms that can cause distortions and biases in academic social media", *Online Information Review*, Vol. ahead-of-print No. ahead-of-print, doi: 10.1108/OIR-01-2021-0042.

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., Broelemann, K., Kasneci, G., Tiropanis, T. and Staab, S. (2020), "Bias in data-driven artificial intelligence systems—an introductory survey", *WIREs Data Mining and Knowledge Discovery*, Vol. 10, p. e1356.

OECD (2019), *Artificial Intelligence in Society*, OECD Publishing, Paris.

Osoba, O.A. and Welser, W. (2017), *An Intelligence in Our Image: the Risks of Bias and Errors in Artificial Intelligence*, Rand Corporation, Santa Monica, CA.

Otterbacher, J., Bates, J. and Clough, P. (2017), "Competent men and warm women: gender stereotypes and backlash in image search results", *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pp. 6620-6631.

Parsheera, S. (2018), "A gendered perspective on artificial intelligence", *10th ITU Academic Conference Kaleidoscope: Machine Learning for a 5G Future, ITU K 2018*, IEE, pp. 1-7.

Petticrew, M. and Roberts, H. (2008), *Systematic Reviews in the Social Sciences: A Practical Guide*, John Wiley & Sons, Malden, MA.

Prates, M.O.R., 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, pp. 6363-6381.

Rahwan, I. (2018), "Society-in-the-loop: programming the algorithmic social contract", *Ethics and Information Technology*, Vol. 20, pp. 5-14.

Raisch, S. and Krakowski, S. (2021), "Artificial intelligence and management: the automation–augmentation paradox", *Academy of Management Review*, Vol. 46, pp. 192-210.

Raji, I.D. and Buolamwini, J. (2019), "Actionable auditing: investigating the impact of publicly naming biased performance results of commercial ai products", *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, Hawaii, pp. 429-435.

Raji, I.D., Smart, A., White, R., N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D. and Barnes, P. (2020), "Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing", *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 33-44.

Roopaei, M., Horst, J., Klaas, E., Foster, G., Salmon-Stephens, T.J. and Grunow, J. (2021), "Women in AI: barriers and solutions", *IEEE World AI IoT Congress (AIIoT)*, 2021, IEEE, pp. 0497-0503.

Saka, E. (2020), "Big data and gender-biased algorithms", in Bachmann, I., Cardo, V., Moorti, S. and Scarcelli, C.M. (Eds), *The International Encyclopedia of Gender, Media, and Communication*, John Wiley & Sons.

Selbst, A.D., Boyd, D., Friedler, S.A., Venkatasubramanian, S. and Vertesi, J. (2019), "Fairness and abstraction in sociotechnical systems", *Proceedings of the Conference on Fairness, Accountability, and Transparency*, Atlanta, Association for Computing Machinery, pp. 59-68.

Shamseer, L., Moher, D., Clarke, M., Ghersi, D., Liberati, A., Petticrew, M., Shekelle, P. and Stewart, L.A. (2015), "Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation", *BMJ*, Vol. 349, jan2015.

Shekhawat, N., Chauhan, A. and Muthiah, S.B. (2019), "Algorithmic privacy and gender bias issues in Google ad settings", *Proceedings of the 10th ACM Conference on Web Science*, pp. 281-285.

Singh, V.K., Chayko, M., Inamdar, R. and Floegel, D. (2020), "Female librarians and male computer programmers? Gender bias in occupational images on digital media platforms", *Journal of the Association for Information Science and Technology*, Vol. 71, pp. 1281-1294.

Smith, R. and Rustagi, I. (2021), "When good algorithms go sexist: why and how to advance AI gender equity", *Stanford Social Innovation Review, SSIR*.

Springer, A., Garcia-Gathright, J. and Cramer, H. (2018), "Assessing and addressing algorithmic bias-but before we get there", *2018 AAAI Spring Symposium Series*, San Francisco, pp. 450-454.

Thelwall, M. (2018), "Gender bias in machine learning for sentiment analysis", *Online Information Review*, Vol. 42 No. 3, pp. 343-354, doi: 10.1108/OIR-05-2017-0153.

Tranfield, D., Denyer, D. and Smart, P. (2003), "Towards a methodology for developing evidence-informed management knowledge by means of systematic review", *British Journal of Management*, Vol. 14, pp. 207-222.

UNESCO (2020), *Artificial Intelligence and Gender Equality*, UNESCO, Paris.

Varghese, S. (2019), Ruha Benjamin: 'we definitely can't wait for silicon valley to become more diverse', *The Guardian*, 29 June.

Veale, M. and Binns, R. (2017), "Fairer machine learning in the real world: mitigating discrimination without collecting sensitive data", *Big Data and Society*, Vol. 4.

Wang, A., Ramaswamy, V.V. and Russakovsky, O. (2022), "Towards intersectionality in machine learning: including more identities, handling underrepresentation, and performing evaluation", *arXiv Preprint*, arXiv:2205.04610.

Wellner, G.P. (2020), "When AI is gender-biased: the effects of biased AI on the everyday experiences of women", *Humana Mente*, Vol. 13, pp. 127-150.

Wellner, G. and Rothman, T. (2020), "Feminist AI: can we expect our AI systems to become feminist?", *Philosophy and Technology*, Vol. 33, pp. 191-205.

West, S.M., Whittaker, M. and Crawford, K. (2019), *Discriminating Systems: Gender, Race and Power in AI*, AI Now Institute.

Weyerer, J.C. and Langer, P.F. (2019), "Garbage in, garbage out: the vicious cycle of AI-based discrimination in the public sector", *Proceedings of the 20th Annual International Conference on Digital Government Research*, pp. 509-511.

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S.M., Richardson, R., Schultz, J. and Schwartz, O. (2018), "AI now report 2018", AI Now Institute at New York University, New York.

Wieringa, M. (2020), "What to account for when accounting for algorithms: a systematic literature review on algorithmic accountability", *Conference on Fairness, Accountability, and Transparency*, Barcelona, Spain. New York, ACM, pp. 1-18.

World Economic Forum (2018), "How to prevent discriminatory outcomes in machine learning", Global Future Council on Human Rights 2016–2018, Switzerland.

Yapo, A. and Weiss, J. (2018), "Ethical implications of bias in machine learning", *Proceedings of the 51st Hawaii International Conference on System Sciences*.

Young, E., Wajcman, J. and Sprejer, L. (2021), "Where are the women? Mapping the gender job gap in AI", Policy Briefing: Full Report The Alan Turing Institute, London, UK.

Corresponding author
Debbie Ellis can be contacted at: vigard@ukzn.ac.za