Gender bias in AI-based decision-making systems: a systematic literature review

Ayesha Nadeem
University of Technology Sydney, Australia
ayesha.nadeem@student.uts.edu.au

Olivera Marjanovic
University of Technology Sydney, Australia

Babak Abedin
Macquaire University, Sydney, Australia

Abstract

The related literature and industry press suggest that artificial intelligence (AI)-based decision-making systems may be biased towards gender, which in turn impacts individuals and societies. The information system (IS) field has recognised the rich contribution of AI-based outcomes and their effects; however, there is a lack of IS research on the management of gender bias in AI-based decision-making systems and its adverse effects. Hence, the rising concern about gender bias in AI-based decision-making systems is gaining attention. In particular, there is a need for a better understanding of contributing factors and effective approaches to mitigating gender bias in AI-based decision-making systems. Therefore, this study contributes to the existing literature by conducting a Systematic Literature Review (SLR) of the extant literature and presenting a theoretical framework for the management of gender bias in AI-based decision-making systems. The SLR results indicate that the research on gender bias in AI-based decision-making systems is not yet well established, highlighting the great potential for future IS research in this area, as articulated in the paper. Based on this review, we conceptualise gender bias in AI-based decision-making systems as a socio-technical problem and propose a theoretical framework that offers a combination of technological, organisational, and societal approaches as well as four propositions to possibly mitigate the biased effects. Lastly, this paper considers future research on the management of gender bias in AI-based decision-making systems in the organisational context.

Keywords: Artificial Intelligence, Fairness, Gender Bias.

1 Introduction

Artificial intelligence (AI)-based decision-making systems are now used in various industry sectors at an increasing rate and continue to penetrate all aspects of our daily lives. The current literature offers many examples of AI-based decision-making systems' benefits. For instance, these systems have the potential to improve organisational operations and decision-making (Kordzadeh & Ghasemaghaei, 2021). Also, recent research indicates that there has been an increased interest in AI-based decision-making systems during the recent COVID-19 pandemic, because of reduced face-to-face human interaction and increased use of automation (Collins et al., 2021). This in turn has further accelerated the use of these systems in various industries.

However, while AI-based decision-making systems may offer solutions to various problems faced in different disciplines, they may simultaneously create unintended harmful effects, including gender-biased outcomes affecting individuals or minorities of a certain race, gender, or colour (Ntoutsi et al., 2019; Eubanks, 2018; Caplan et al., 2018; Benjamin, 2019; West et al. 2019; UNESCO 2020). Worryingly, AI-based decision-making systems are increasingly used to screen job applications, determine outcomes of loan applications, calculate insurance premiums and benefits, determine access to social services and more (Mehrabi et al., 2019; Caplan et al., 2018; Marjanovic et al., 2021). The instances of gender-biased AI-based decision-making systems have already been reported in the scientific literature and popular press. For instance, Facebook's job ads, highly favoured males for STEM (science, technology, engineering, and mathematics) jobs (Lambrecht & Tucker, 2019), and credit loan applications. Also, Amazon discontinued using an AI-based decision-making system for recruitment, which resulted in gender-biased outcomes (Dastin, 2018; Bolukbasi et al., 2016; Kordzadeh, & Ghasemaghaei, 2022). The AI gender-biased outcomes was due to the lack of female applicant data incorporated in the training datasets.

The concern regarding gender bias in AI-based decision-making systems has also been raised by governments and research organizations (Parikh et al., 2019; Feast, 2019; Parsheera, 2018; Agarwal, 2020). The harmful effects of these systems go beyond individuals and are reported to affect families, communities, and society at large (Altman et al., 2018). Therefore, it is important to scrutinise AI-based decision-making systems for gender bias in order to ensure fairness in its outcomes, which is one of the fundamental principles of AI ethics (Mehrabi et al., 2019; Jobin et al., 2019). A greater understanding of this type of bias will also help organisations to make conscious strategic choices (Marabelli et al., 2021).

Given the above, this paper aims to contribute to the emerging body of IS literature on the unintended harmful effects of AI by focusing on gender bias in AI-based decision-making systems. The objectives of this study are (i) to identify and examine the characteristics of gender bias in AI-based decision-making systems, (ii) to investigate the role of relevant contributing factors behind gender bias in AI-based decision-making systems and the reported approaches to mitigation of gender bias in AI-based decision-making systems, and (iii) to propose a theoretical framework for the management of gender bias in AI-based decision-making systems.

This paper is organised as follows: section 2 introduces the foundational concepts and further elaborates on the significance of this research; section 3 describes the adopted research method that is a systematic literature review process, along with the search criteria, selection of articles and analysis, and coding process; section 4 presents the findings of this research; section 5 discusses the proposed framework along with considerations for future research; section 6 offers the conclusion and discusses the study limitations.

2 Foundational Concepts

AI-based IT systems transform IT systems from just representing reality to also actively participating in it and influencing it, and thereby these systems are explicitly demonstrating digital agency (Baskerville et al., 2020; Niehaus, & Wiesche, 2021). In this research we follow Markus (2017) and refer to a particular type of AI-based IT systems which automate algorithmic decision-making based on computational models and among others natural language processing capabilities as AI-based decision-making systems where "Automated decision-making is the process of making a decision by automated means without any human involvement. These decisions can be based on factual data, as well as on digitally created profiles or inferred data. Examples of this include: an online decision to award a loan; and an aptitude test used for recruitment which uses pre-programmed algorithms and criteria" (ICO, 2018, p. 5).

AI-based decision-making systems are now underpinning the digital economy; at the same time, they are also criticised regarding their fairness, accountability, and transparency (Feuerriegel et al., 2020). Consequently, there has been an outburst of research on fairness in AI-based decision-making systems in recent years (Feuerriegel et al., 2020; Bellamy et al., 2018; Zhong, 2018; Leavy, 2018; Jobin et al., 2019). Moreover, considerations of fairness in AI-based decision-making systems in organisations are still lagging, including fair practices within systems, people, and processes (Feuerriegel et al., 2020). Hence, IS researchers and practitioners have been encouraged to work and collaborate towards 'fair AI' (Feuerriegel et al., 2020). This also includes increasing concerns about, and reconsideration of the current approaches to bring fairness to AI-based decision-making systems (Ntoutsi et al., 2019; Feuerriegel et al., 2020; Kordzadeh, & Ghasemaghaei, 2021).

Dwivedi et al. (2019) argue that it is imperative to study fairness in AI-based decision-making systems as they are limited to industrial applications but have entered our lives on a daily basis. Yet, the notion of 'fairness' remains unclear. For example, there are various, even mutually incompatible, definitions of fairness proposed by computer science researchers, with system and software developers unable to resolve these differences (Teodorescu et al., 2021). At the same time, there are long-standing discussions on fairness within philosophical and theological literature for centuries, often in connection with justice (Feuerriegel et al., 2020). In the absence of any well-established definition of fairness, in this paper, we draw from the previous work by Merhrabi et al. (2019) who consider fairness as the elimination of any prejudice or favouritism behaviour towards a certain group or individuals. According to Hayes et al., (2020), fairness prevents any action or policy that perpetuates discrimination or unequal treatment. Hence, fairness refers to treating others the way one wants to be treated (Teodorescu et al., 2021). An example of unfairness could be the act of disqualifying individuals who want to improve their financial conditions by rejecting their loan applications or job applications based on their gender, ethnicity, or the neighbourhood they live in (Feuerriegel et al., 2020).

While the concept of fairness is very broad, gender-related fairness is considered an essential aspect of fairness. Gender bias, as defined by Masiero and Aaltonen (2020, p.1), is 'the systemic, unfair difference in a way men and women are treated in a particular domain'. The related literature now provides strong evidence about gender bias in some AI-based decision-making systems (Agarwal, 2020; Altman et al., 2018; Bolukbasi et al., 2016; Canetti et al., 2019; Crawford, 2016; Dwivedi et al., 2019; Galleno et al., 2019; Lambrecht, & Tucker 2019; Mehrabi et al., 2019; Nadeem et al., 2020; Trewin et al., 2019). However, the research on gender-related biases in AI-based decision–making systems in IS is still emerging (Jobin et al., 2019; Marabelli et al., 2021) there are still research gaps regarding our understanding of gender bias in AI-based decision-making systems, particularly what causes gender bias in AI-based decision-making systems, the mitigation of this bias and possible prevention (Leavy, 2018). Additionally, there is a significant lack of research on how to manage bias in AI-based decision-making systems, including its harmful implications (Berente et al., 2019; Feuerriegel et al., 2020; Kordzadeh, & Ghasemaghaei, 2021). Therefore, approaches to prevent, tackle and mitigate gender bias in AI-based decision-making systems are of high priority.

3 Methodology

To achieve the objectives of this research, we conducted a systematic literature review (SLR) of the related literature in IS and beyond. This is an appropriate research method as the research phenomenon is still emerging and a SLR can be used to it systematically summarise and investigate previous findings (Cao et al., 2015; Webster & Watson, 2002). The outcomes of a SLR can further be used as a valued reference for future research (Kitchenham et al., 2011; Petersen et al., 2015, Pare` et al., 2015). As Borges et al. (2021) observe, the analysis of articles selected through SLR yields a rich picture of various characteristics. Also, systematic reviews allow researchers to examine the scope and range of research activities in a given domain by focusing on the breadth of the literature covered (Pare` et al., 2015).

Our study adopts the SLR approach introduced by Bandara et al. (2011) and Wolfswinkel et al. (2013). The approach enables researchers to conduct a conceptualised analysis of the literature and identify the key themes (Wolfswinkel et al., 2013). The process of selection and identification of relevant articles was carried out using a rigorous method, as shown in Figure 1.

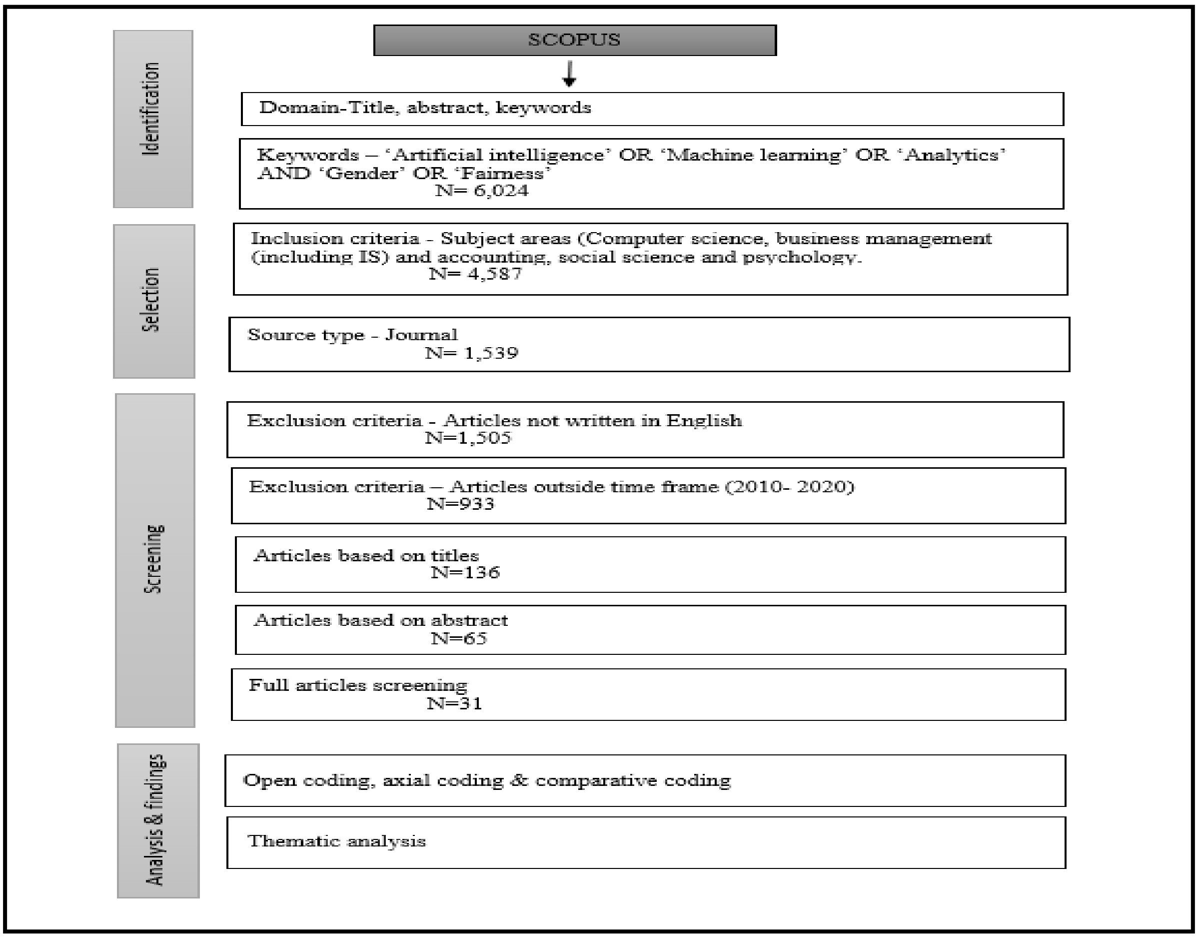

Figure 1. Selection of articles

3.1 Search criteria

For this study, we used Scopus, which is one of the largest databases for journals and books (Collins et al., 2021), and which has also been used by numerous AI researchers (Borges et al., 2021).

The first step of our SLR included a thorough investigation of the appropriate keywords' selection. We initially performed a generic and multidisciplinary literature review (Nadeem et al., 2020) for an iterative process of refining and selecting identified relevant keywords. The keywords were then further reviewed and selected based on the scope of this study. Subsequently, we conducted a keyword search for the period from 2010 to 2020 to capture published research on gender bias in AI-based decision-making systems. The time frame of 10 years includes all relevant studies in high-quality journals is considered a reasonable and recommended time frame (Borges et al., 2021). We selected computer science and business management (including IS) as the subject areas. We also included social sciences and psychology to cover the social and behavioural aspects of gender bias.

3.2 Selection of the articles

The identification of articles in Scopus started after the selection of the keywords, i.e., artificial intelligence, machine learning, analytics, gender, fairness. A total of 6,024 articles were captured through the selected keywords. To further filter the relevant articles, we applied inclusion criteria and source type (see Figure 1). This was followed by exclusion criteria and time frame; articles that were outside the time frame (i.e., 2010 – 2020) and that were not written in English were excluded from the final set of articles. Then, we started by reading the titles and abstracts of the identified articles. After selecting the articles on the basis of their titles and abstracts, we thoroughly read the full text of the articles. In this step, we considered only those articles that were directly dealing with gender bias in AI-based decision-making systems. Hence, we excluded all papers that were outside the scope of this research, and ultimately 31 papers were selected that were relevant to our research scope.

3.3 Analysis of the selected articles

The analysis of the final set of 31 papers was carried out by in-depth reading of the articles. The relevant concepts and themes were identified by open coding, axial coding, and comparative analysis, as suggested by Wolfswinkel et al., (2013), and through thematic analysis (Pare` et al., 2015). The themes were initially coded by the first author independently and inductively, and then they were scrutinised by the other two authors for authentic and unbiased themes and outcomes. In the coding process, all codes with similar themes were integrated into one concept; f. ex. for the characteristics of gender bias in AI-based decision-making systems, the codes 'societal gender prejudices' and 'discrimination' were merged into 'prejudices in society' and later integrated into a concept 'societal' as shown in the appendix in Table 2. Similarly, when coding the approaches for mitigating gender bias in AI-based decision-making systems, the codes 'bias aware collection of datasets', 'preparation of fair data sets' and 'removing proxies of protected attributes in data sets' were merged into the theme 'collection and preparation of dataset', which was later integrated into the concept 'AI technological approaches' as shown in the appendix in Table 4.

4 Findings Of The Systematic Literature Review

In the following, we present the key findings from the SLR, which identify and categorise the insights about the manifestations of gender bias in AI-based decision-making systems and the contributing factors, as well as possible approaches to mitigate it. In doing so, we contribute to the emerging body of IS literature on the potentially harmful effects of AI and their mitigation.

To reach an in-depth understanding of this area of research, we first identified and noted the type of published articles (i.e., conceptual research, literature review, design science, empirical, survey or case-based research), their use of theory and their focus as presented in Table 1 in the appendix.

Our analysis shows that the majority of the reviewed articles have been conceptual papers. There is a lack of empirical and detailed literature review papers. Moreover, the publication trend, as depicted by Figure 2, indicates that the number of relevant publications started to grow in 2017 and then increased quite considerably in 2020, which confirms that this is a fast-growing research area. Further, our analysis discovers that although the topics of fairness and gender bias in AI have been discussed in the broader literature, the IS field is yet to pay more attention to this important topic. Our findings also indicate that the research on gender bias in AI is not yet well established, which highlights a great potential for future research in the IS field and beyond.

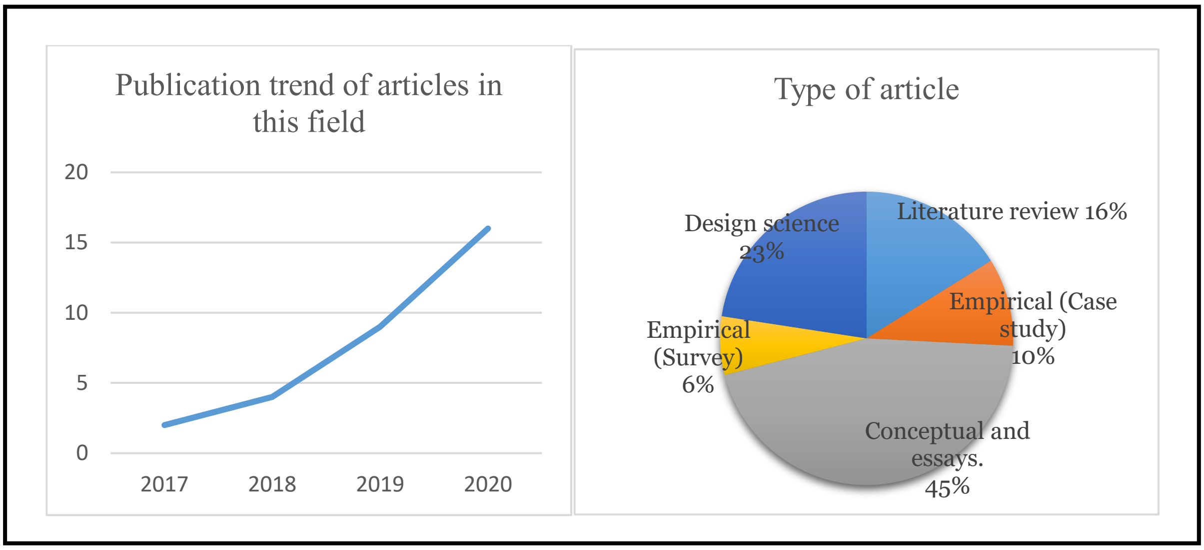

Figure 2. Publication trend of reviewed articles and type of selected articles

4.1 Characteristics and Contributing Factors of Gender Bias in AI-Based Decision-Making Systems

We adopt the term 'characteristics' to indicate various domains of gender bias in AI-based decision-making systems and observe three main characteristics of gender bias in AI-based decision-making systems: design and implementation, institutional and societal.

The identified characteristics are intertwined (see Figure 3). The long-standing societal inequalities and discriminatory norms propagate to organisational culture, thus affecting institutional practices and are manifest in the design and implementation of AI-based decision-making systems.

When considering the contributing factors of gender bias in AI-based decision-making systems, we established six main themes relating to: gender stereotyping, biased training datasets, lack of gender diversity in AI development teams, AI amplifies existing bias, contextual and other factors and lack of AI regulations, (see also depicted by Figure 3). These contributing factors are rooted within the characteristics of gender bias in AI-based decision-making systems. Table 2 and 3 in the appendix include the sources, description and coding process of the characteristics and contributing factors of gender bias in AI-based decision-making systems and are further discussed in the next subsection.

4.1.1 Design and Implementation Characteristics

The design of AI-based decision-making systems including any possible flaw in this phase can have an impact on the implementation and use of these systems (Marabelli et al., 2021). A major challenge and reason for such flaws in the design of such systems is the misrepresentation of the datasets, i.e., ones that are either biased, incomplete, or incorrect (Marabelli et al., 2021). According to Hayes et al. (2020), societal gender inequalities are incorporated in the AI algorithms datasets due to unfair representation of datasets (i.e. over, under or misrepresentation of certain groups) and lack of gender diversity in the design of AI-based decision-making systems that create 'blind spots' (Johnson, 2019; Lee, 2018; Wang, 2020; Martinez & Fernandez, 2020; Clifton et al., 2020). Based on our literature review, the design and implementation characteristics are found to have the following contributing factors.

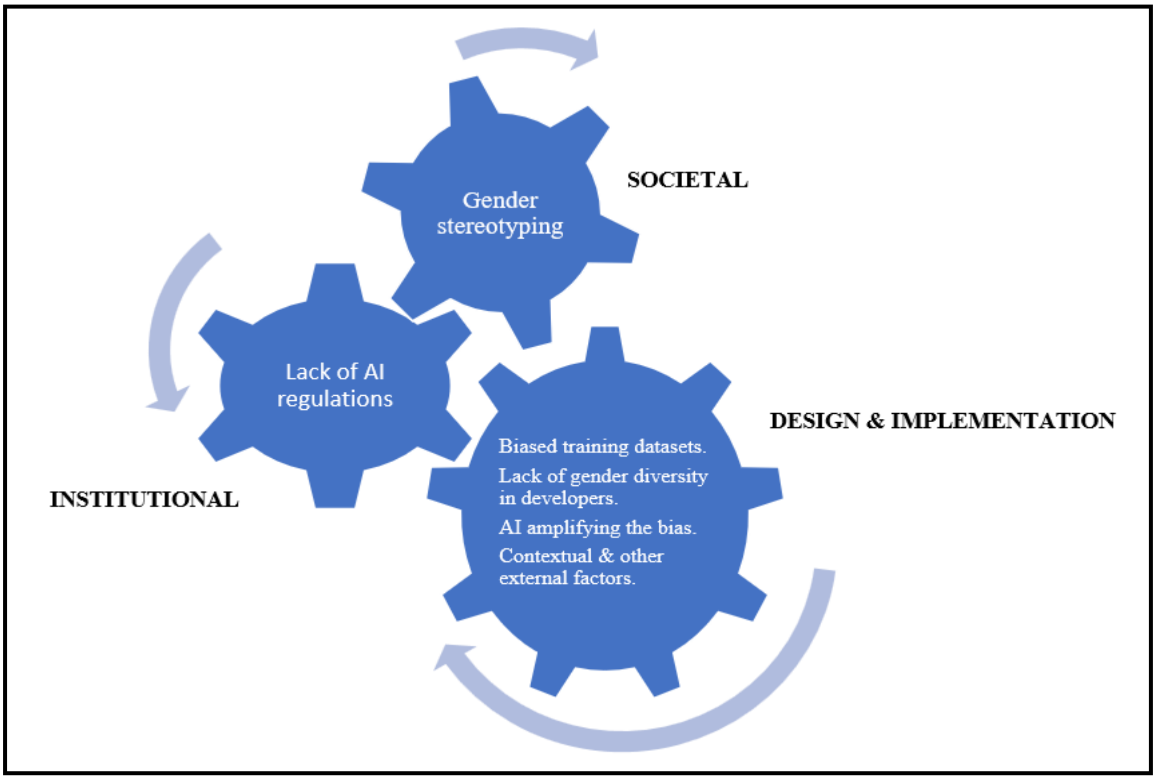

Figure 3. Characteristics and contributing factors of gender bias in AI-based decision-making systems

Biased training datasets are patterns of unfairness in datasets (Veale & Binns, 2017) that are often based on the under- or over-representation of social groups and that convert the computational training processes to a biased discriminative decision (Ntoutsi et al., 2019). Additionally, the correlation of data of sensitive variables and features (i.e., proxy variables) makes its way into AI algorithms and modelling and results in biased outcomes. Past literature has noted that proxy variables discriminate against certain groups (e.g., salary serving as a proxy for gender and zip code serving as a proxy for background) (Feuerriegel et al., 2020; Ahn & Lin, 2020; Martinez & Fernandez, 2020). Further, word embedding not only preserves the statistical relationship present in the training data but also places co-occurring words close to each other, such as man is to king and is woman to queen (Mikolov et al., 2013; Brunet et al., 2019) and man is to computer programmer as woman is to homemaker (Bolukbasi et al., 2016; Brunet et al., 2019). Also, gender stereotyping is predominant across different word embedding practices (see subsection 4.1.3) and as such not an artefact of a particular word training corpus or methodology (Bolukbasi et al., 2016). For instance, female pronouns such as her or she, and the word woman are closely associated with family and arts, while the term male is largely associated with career, intelligence and maths (Brunet et al., 2019).

Lack of gender diversity in AI development teams as shown in the literature confirms that the gender disparity in systems and software developers and data miners may result in gender bias during the training phase of the algorithms of AI-based decision-making systems (Clifton et al., 2020; Johnson, 2019; Lee, 2018; Martinez & Fernandez, 2020; Wang, 2020). As stated by Feuerriegel et al. (2020), the lack of gender diversity in AI developers and other workers in Science, Technology, Engineering and Mathematics (STEM) careers is reflective of a male-dominated, homogeneous IT industry, which may lead to a lack of diversity of mindsets in development teams (Johnson, 2019; Lee, 2018; Wang, 2020) that develop AI-based decision-making systems. In turn, this may reinforce the dominance of one gender (male) and control over algorithms and decisions, yielding gender-biased outcomes. For instance, facial recognition software being used in USA in 2015 was unable to handle diversity well (Daugherty et al., 2018; Otterloo, 2019), and a huge amount of research has examined the difference in treatments of men and women in US criminal justice, with women being more likely to be arrested and sentenced than men (Kulik et al., 1996). Further, Lambrecht & Tucker (2019) investigated how advertisements promoting job opportunities in STEM are viewed by more men than women, which eventually results in fewer women's applications for STEM jobs.

The lack of human feedback and 'humans-in-the-loop' in AI-based decision-making systems may also **amplify existing bias** (Johnson, 2019; Miron et al., 2020) as well as the choice and application of certain modelling approaches during the training of the algorithms (Chen, et al., 2019, Feuerriegel et al., 2020;). Once inscribed into the algorithmic training datasets, this bias is perpetuated due to the systems' self-training. Research found that some AI-based decision-making systems work better for certain groups of people over time, thus perpetuating inequalities in society by learning through biased outcomes. For instance, Teodorescu et al. (2021) uncovered that gender disparity was perpetuated by Facebook advertisements' targeted job posting algorithms in which female applicants failed to see the job advertisements of companies that predominately hired male applicants. Hence, if such algorithms are opaque - and complex - (Miron et al., 2020) they may re-enforce their creators', programmers', developers', designers', software engineers' and data miners' bias (Hayes et al., 2020; Miron et al., 2020; Ntoutsi et al., 2019; Wang, 2020), further yielding gender bias in AI-based decision-making systems.

The lack of proper testing of an algorithm for specific contexts may lead to decisions that disadvantage certain social groups in society (Qureshi et al., 2020). In the context of AI-based decision-making systems, important **contextual and external factors** are often left unnoticed (Marabelli et al., 2021). Such contextual and external factors like third parties collecting the data, in the process might omit some important variables (Johnson 2019; Ntoutsi et al., 2019) and the lack of proper testing of an algorithm for specific contexts may lead to decisions that disadvantage certain social groups in society (Qureshi et al., 2020).

4.1.2 Institutional Characteristics

Gender bias in AI-based decision-making systems is also manifested within institutions, with gender-biased decisions reported to be influenced by a broader societal context (see also subsection 4.1.3). AI-based decision-making systems both reflect and amplify the existing societal bias.

In our literature review, we identified that a **lack of AI regulations** is a significant contributing factor to the institutional characteristics impacting on such systems. Due to conscious or unconscious categorization between sociocultural groups, some institutions operate in ways that might disadvantage some minorities or social groups because of socioeconomic factors (Costa & Ribas, 2019; Ntoutsi et al., 2019) (see also subsection 4.1.3). Further, there is limited development in regulations toward addressing gender bias in AI-based decision-making systems (Johnson, 2019). Despite the European Union's (2021) recently presented proposal for firmer AI regulations (Marabelli et al., 2021), there is a lack of more-precise AI guidelines for developers and institutions concerning fair AI in particular regarding data protection and data quality (Johnson, 2019; Ntoutsi et al., 2019; Wang, 2020). Thus, societal gender stereotyping and discrimination are amplified through institutional characteristics. This underlines the need and urgency for AI regulations and policy intervention for fairer AI (Hoffmann, 2019; Lee, 2018; Ntoutsi et al., 2019).

4.1.3 Societal Characteristics

The long-standing inequalities in society, i.e., gender stereotyping (Johnson 2019; Ntoutsi et al., 2019) leading to preferential treatment towards masculinity are often reflected in AI algorithms.

These concepts normally 'sneak in' the datasets through the misguided conduct of 'bad actors' (Hoffmann, 2019). Hence, they connect the already-existing concept of gender stereotyping in society to gender bias in AI-based decision-making systems (Cirillo et al., 2020; Clifton et al., 2020; Johnson, 2019; Martinez & Fernandez, 2020; Noriega, 2020; Ntoutsi et al., 2019; Sun et al., 2019; Wang, 2020

Based on our literature review, we found that **gender stereotyping** as a contributing factor to gender bias in AI-based decision-making systems occurs in societies where historical biases and norms are followed not because of conscious discrimination but rather because the majority following the pre-existing customs presents a culture that promotes masculinity and exclusivity (Johnson, 2019; Ntoutsi et al., 2019), this biased behaviour is inscribed in AI systems; therefore, AI-based decision-making systems reflect human biases toward people from a certain background, race, or gender. Also, certain upstream social norms are followed blindly because of their easy acceptance in society (Ntoutsi et al., 2019). For instance, certain professions are associated with males e.g., doctors, engineers, and scientists, while professions like nursing and secretary work are associated with females. The wording and association of certain professions with certain genders sow unequal division and discrimination in society.

Related to gender stereotyping, socioeconomic factors also impact and amplify gender bias in AI-based decision-making systems i.e., socio-economic factors based on social standing (e.g., neighbourhood, zip code and location) result in incorrect assumptions of an individual. People with lower socio-economic backgrounds may be disadvantaged in a society due to their social status and standing. This biased behaviour may be reflected in AI-based decision-making systems because of the unconscious bias of those who develop these systems (Clifton et al., 2020; Wang, 2020) For instance, the Amazon delivery system excluded certain socio-economic neighbourhoods due to socioeconomic stereotyping in the society impacting AI-based decision-making systems and their outcomes (Dastin, 2018).

4.2 Approaches to Mitigating Gender Bias in AI-Based Decision-Making Systems

Based on the reviewed literature, we observe proposals for four main approaches to possibly mitigate gender bias in AI-based decision-making systems, which are: AI technology-related approaches, fair AI management approaches, AI governance and regulatory approaches, and societal and community-focused approaches. Table 4 in the appendix presents the sources, description, and coding process, which resulted in the identified approaches.

4.2.1 AI technology-related approaches

Clifton et al. (2020) propose a strategy of capturing data from all vulnerable, gender-diverse groups of society and adding multi-dimensional datasets during the design of the decision-making algorithms, which in their view can neutralise inappropriate and/or unfair datasets. Similarly, Hayes et al. (2020) argue that a fair representation of the population in data sets will result in fair AI outcomes.

Researchers also call for AI-based decision-making algorithms to be designed and programmed in such a way that they do not replicate prejudices and gender bias while analysing and interpreting the data (Johnson 2019; Ntoutsi et al., 2019). If the implementation context does not match training datasets, the resulting AI-based decision-making systems are unlikely to perform well, i.e., lead to biased outcomes (Hardt & Price, 2016). Thus, testing the algorithms for a specific application increases the accountability and bias detection procedures (Ahn & Lin, 2020; Arrieta et al., 2020; Bellamy et al., 2018; Berk et al., 2018; Feuerriegel et al. 2020; Grari et al., 2020; Johnson, 2019; Lambrecht, & Tucker, 2019; Martin, 2019; Miron et al., 2020; Ntoutsi et al., 2019; Thelwall, 2017; Veale & Binns, 2017).

Context-specific decision-making algorithms are put forward as being more effective; however, they require continuous re-design as per the specific contextual conditions (Marabelli et al., 2021). Appropriate choices for AI-based decision-making systems need to be made, depending on the context in which they are being used (Marabelli et al., 2021). Moreover, paying additional attention to the context of the AI-based decision-making systems, data and people involved can effectively decrease their discriminatory outcomes (Marabelli et al., 2021).

4.2.2 Fair AI management approaches

Hayes et al. (2020) found that, due to unethical data practices, misreporting of data and other misconducts related to data collection and preparation as well as the ability of AI decision-making algorithms to self-learn from the so-called *pernicious feedback* of their own, biased decisions worsen the AI outcomes. To address this concern, researchers (Johnson, 2019; Miron et al., 2020) suggest incorporating testing and auditing the AI-based decision-making algorithms into the design and implementation phase. This could involve external auditors or internal compliance auditors (Martinez & Fernandez, 2020; Kyriazanos et al., 2019). For example, AI experts can maintain regular testing and verification of AI-based decision-making systems and can also use interpretation tools to diagnose potential problems and challenges (Wu et al., 2019).

Other recommended strategies focus on human-decision makers. If given authority, humans actively involved in the ultimate decision-making can effectively adjust the outcomes provided by the technical components of AI-based decision-making systems (Hayes et al., 2020). Further inclusiveness and diversity training to decision-makers is also suggested as an approach to avoid unconscious bias and, most importantly, understand and identify of when to intervene in the AI proposed decision (Hayes et al., 2020). Likewise, institution-wide education that involves principles of ethics, such as promoting ethical education for every stakeholder involved in AI practices (Martin, 2019; Noriega, 2020), is reported to assist in detecting gender-biased outcomes. Other education-based approaches include professional certifications and courses focused on building awareness of gender bias in AI-based decision-making systems (Martin, 2019).

In addition, increased and enhanced AI corporate governance regarding gender inclusion in the development of AI technologies is suggested as a strategy to introduce diverse perspectives (Costa & Ribas, 2019; Johnson, 2019; Lee, 2018; Ntoutsi et al., 2019) and to ensure that gender bias is addressed in AI (Ibrahim et al., 2020). Enhanced gender diversity and inclusion in the technology sector, especially in development for AI-based decision-making systems (Lambrecht & Tucker, 2019), is proposed to avoid homogeneous and predominantly male-dominated leaderships and decisions (Johnson, 2019). When applied to AI, the inclusion of a more diverse IT workforce in the design and implementation of algorithms and diversity of thoughts in AI development is reported to bring a multicultural perspective to AI design, which in turn might mitigate gender bias (Ibrahim et al., 2020; Wu et al., 2019).

Organisations also need to develop fair and ethical internal structures, corporate strategies, and governance to manage gender imbalance; gender diversity among board members, management, senior developers, and general leadership encourages people from diverse backgrounds, and offers pathway towards mitigating gender bias (Johnson, 2019).

4.2.3 AI governance and regulatory approaches

The beforementioned recent proposal by the European Union (2021) highlights the significance and urgency of creating AI regulations for dealing with humans. Introducing rules and policies governing AI-based decision-making systems ensures better efficiency in the resulting decisions (Marabelli et al., 2021).

Hayes et al. (2020) argue that institutional AI regulations should be developed and implemented to increase transparency and accountability in AI-based decision-making systems. At the same time, AI Algorithms should not be designed in a way that precludes individuals from taking responsibility (Martin, 2019). Similarly, researchers argue that users who are affected by AI decisions should have the right to know and comprehend the reasons behind those decisions and share their feedback on them (Wu et al., 2019). Having regulations relating to formal verification, adhering to AI ethical values and testing AI-based decision-making systems in place, is envisaged to result in fair outcomes (Lee, 2018; Wu et al., 2019). This includes datasets purchased from third parties that need to be properly analysed for the particular context in which they will be used (Johnson, 2019), as well as enhanced ethical AI standards by government and regulatory organisations about data collection and selection (Cirillo et al., 2020).

AI governance across interdisciplinary and multinational collaborations is suggested to establish a census on AI principles, which in turn enhances the general practice of responsible

AI conduct (Wu et al., 2019). Recognising the need for knowledge sharing, researchers propose collaborative ethical AI online platforms for all stakeholders, which would permit demographically diverse organisations to collaborate and share knowledge regarding the appropriate and practical strategies to promote fairness in AI systems (Soleimani et al., 2021; Veale & Binns, 2017).

4.2.4 Societal and community-focused approaches

Hayes et al. (2020) and Prates et al. (2019) propose to encourage social interventions such as enhancing professional education and training on gender diversity in the community to boost diversity and inclusiveness. Public policies to protect the personal data of users are also proposed as a possible approach to increase confidence in AI-based decision-making systems, in particular when it comes to sharing personal data for such decision-making process (Clifton et al., 2020). Cirillo et al. (2020) propose an 'ecosystem of trust' by government or policymakers to ensure that systems comply with the fundamental rules that protect both human and consumer rights, particularly in AI-based decision-making systems.

5 Discussion and Contributions

The findings of this study contribute to an improved understanding of the state of gender bias in AI-based decision-making systems. So far, there has been ample work on exploring and reducing the bias in AI-based decision-making systems through technical approaches. Research communities such as FAT machine learning (fairness, accountability, and transparency in machine learning) have emphasised bringing fairness to AI algorithms through programming and mathematical modelling (Veale & Binns, 2017). Consequently, many AI researchers see gender bias in AI-based decision-making systems as a technological problem (Ahn & Lin, 2020; Arrieta et al., 2020; Bellamy et al., 2018; Grari et al., 2020; Lee 2018; Miron et al., 2020; Ntoutsi et al., 2019; Veale & Binns, 2017). However, the roots of gender bias in AI-based decision-making systems are not technological, and thus, technical solutions might not suffice (Nadeem et al., 2021). Information systems research lags behind in addressing the behavioural, organisational, and social implications, antecedents, and consequences of this problem, despite the fact that computational scientists have developed mathematical techniques to detect and mitigate biases in algorithms (Kordzadeh & Ghasemaghaei, 2021; Sarker et al., 2019).

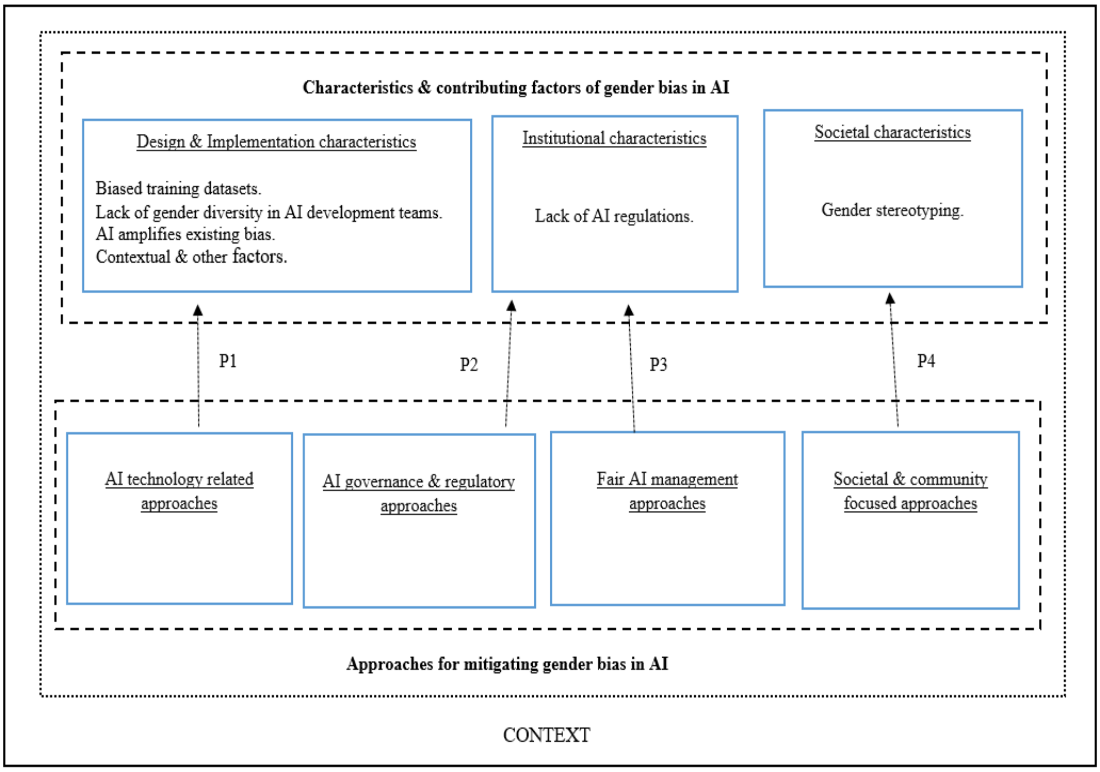

Figure 4. Proposed framework for the management of gender bias in AI-based decision-making systems

The bias in AI-based decision-making systems cannot be corrected by merely fixing the decision-making algorithms; 'this is not an algorithmic problem', as stated by Teodorescu et al. (2021). AI-based decision-making systems are multidisciplinary phenomena that call for the collaboration of experts representing technological, organizational, and human perspectives (Marabelli et al., 2021). Moreover, bias in decision-making algorithms is socio-technical in nature, and thus social implications of this phenomenon should be at the centre of its examination and potential solution (Kordzadeh & Ghasemaghaei, 2021).

Consequently, we conceptualise gender bias in AI-based decision-making systems as multi-layered, multidimensional, and socio-technical with the systems' development and implementation requiring a combination and integration of technical, organizational, and societal approaches.

For this purpose, we systematically reviewed the existing literature and build in particular on the previous work from Marabelli et al. (2021) and Kordzadeh & Ghasemaghaei (2021) to advance the conversation about possible technological, organizational, human and societal mitigating approaches. As a result, we propose a theoretical framework for the management of gender bias in AI-based decision-making systems (see below Figure 4). The proposed theoretical framework synthesizes previously reported contributing factors and approaches and consolidates them in a theoretical framework.

As part of the proposed framework and as a summary of our findings we offer four theoretical propositions for the possible mitigation of gender bias in AI-based decision-making systems.

P1: AI technology-related approaches can mitigate design and implementation-related contributing factors.

Removing proxy variables of protected attributes and ensuring fair datasets from all groups and members of a community i.e., diverse and inclusive datasets, are reported to be effective in mitigating gender bias in the design and implementation of AI systems (Bellamy et al., 2018; Feuerriegel et al., 2020; Grari et al., 2020; Hayes et al., 2020; Miron et al., 2020; Veale & Binns, 2017). Having a document or guideline on fair datasets for developers can support fair outcomes and prevent unfairness in training data by ensuring fair data collection, data preparation and regularising the training data to minimise the unfairness (Bellamy et al., 2018; Ntoutsi et al., 2019). Such measures, which strictly speaking are socio-technical approaches, could be the pairing of data scientists with social scientists to achieve multidisciplinary for the design and implementation and for effectively mitigating gender bias in AI-based decision-making systems (Marabelli et al., 2021). Further, enhanced and constant testing for algorithmic accountability and transparency can improve the understanding and explanation of bias detection of algorithmic models and structures (Ntoutsi et al., 2019).

Hence, we propose that AI technology-related and diversity mitigating approaches can be used to address the design and implementation-related factors that contribute to gender bias in AI-based decision-making systems (Bellamy et al., 2018; Feuerriegel, et al., 2020; Grari et al., 2020; Hayes et al., 2020; Johnson, 2019; Lee, 2018; Miron et al., 2020; Noriega, 2020; Ntoutsi et al., 2019; Veale & Binns, 2017; Wu et al., 2019).

P2: AI governance and regulatory approaches can mitigate institutional-related contributing factors.

AI regulations that enforce to incorporate key ethical standards (Ntoutsi et al., 2019; Wang, 2020), adhering to laws and policies for better AI governance, auditing and gender diversity and inclusiveness in organisations concerning fair AI (Feuerriegel et al., 2020) are all reported to result in mitigating gender bias in AI-based decision-making systems. Hence, we propose that AI governance and regulatory approaches can be used to mitigate the institutional-related contributing factors (Feuerriegel et al., 2020; Ntoutsi et al., 2019; Wang, 2020).

P3: Fair AI management approaches can mitigate institutional-related contributing factors.

AI to be fair by design (Arrieta et al., 2020) practiced in organisations with policies and business models concerning fair AI (Feuerriegel et al., 2020) includes implementing inclusive policies and regulations within the organisations and bringing about algorithmic accountability and transparency (Johnson 2019; Ntoutsi et al., 2019).

Fair AI management mitigation approaches through awareness and promoting policies are reported to ensure having ‘humans in the loop’ which increases the chance of fairness provided by AI-based decision-making systems (Teodorescu et al., 2021). In particular, creating awareness through training, workshops, and seminars at the organisational level regarding gender-biased outcomes of AI-based decision-making systems can encourage AI developers and users of such systems to enforce gender-diverse workplaces and public policies regarding fairness to support demographic and cultural diversity in data that is used by the systems (Lee, 2018).

Hence, we propose that fair AI management approaches can be used to mitigate the institutional-related contributing factors (Hayes et al., 2020; Lee, 2018; Marabelli et al., 2021; Teodorescu et al. 2021; Wu et al., 2019).

P4: Societal and community-focused approaches can mitigate societal-related contributing factors.

Gender bias in society is found to be replicated in emerging technologies, i.e., including AI (Kordzadeh & Ghasemaghaei, 2021). Emphasising social interventions – f. ex., awareness of gender equity and fairness in society through social and educational aspects such as workshops, seminars, etc. – is reported to be effective in mitigating the socially manifested gender bias in society (Hayes et al., 2020; Johnson, 2019). Moreover, certain public policies that protect fundamental rights and societal well-being, if enhanced, bring awareness to human rights and work against gender bias and other discrimination (Clifton et al., 2020; Miron et al., 2020).

Hence, we propose that societal & community-focused approaches can be used to mitigate societal-related contributing factors (Clifton et al., 2020; Hayes et al., 2020; Johnson, 2019; Kordzadeh, & Ghasemaghaei, 2021; Miron et al., 2020).

6 Future research

The offered propositions are not exhaustive. Therefore, further research is needed to develop more propositions which along with the ones we have proposed, must be refined and empirically tested.

Based on the proposed theoretical framework, we suggest future IS research related to the prevention, mitigation, and future theorizing of gender bias in AI-based decision-making systems from an IS perspective. While this study presents proposed approaches for mitigating the contributing factors that generate gender bias in AI-based decision-making systems through systematically reviewing the existing literature, it is important to empirically investigate the proposed approaches. Another interesting opportunity for further research is to study how societal gender bias is manifested in institutional AI practices and vice versa as there is a lack of contextually rich theories in this domain, that examine these practices in broader institutional and regulatory structures (Conboy et al., 2022) will be interesting. In this context it is interesting to investigate how regulations can shape gender bias in AI in an organisational context.

We therefore suggest to further evolve the proposed theoretical framework for the management of gender bias in AI-based decision-making systems through future theoretical and empirical research and contribution. We expect such research to build the foundations for new frameworks that, as our systematic literature review confirms, are very much needed. More research on context-specific AI algorithms will be beneficial. Hence, the proposed framework could be further explored in specific organisational and societal contexts.

7 Concluding Remarks and Study Limitations

Gender-related bias is of vital concern in AI-based decision-making systems that are now used in organisational and societal contexts. Therefore, it is important to unpack the status of gender bias in AI-based decision-making systems in the literature and systematically analyse the findings of such reviews for better understanding and mitigation of possible harmful effects. This paper has contributed to the conversation on gender bias in AI-based decision-making systems by identifying and investigating reported design and implementation, institutional and societal approaches to potentially mitigating gender bias in AI-based decision-making systems.

We conceptualise gender bias in AI-based decision-making systems as a socio-technical problem that affects a variety of stakeholders, including the workforce, and society in general. We identify some key characteristics that manifest gender bias in AI-based decision-making systems along with the associated contributing factors and possible approaches for potential mitigation that we developed based on the existing literature and timely industry examples. Hence, a framework is proposed to guide AI designers, developers, and other stakeholders to ensure the management of AI by mitigating gender bias in AI-based decision-making systems.

Our research findings also suggest that organisations need to be actively engaged in the implementation of ethical and fair AI outcomes. In particular, our findings highlight strategies related to workplace diversity, further education on ethical and fair AI as well as improved transparency and accountability in algorithmics. In addition, training and certification on ethical and fair AI should be considered for new employees and reinforced periodically. Moreover, organisations should have AI governance strategies in place, which should support the prevention, detection, and mitigation of gender bias in their AI-based decision-making systems.

We recognise that our study has several limitations. First, different keywords are likely to result in a different pool of related research work. Moreover, a systematic literature review of a wider group of IS and other journals, along with more databases such as the ISI web of science could be used to obtain more detailed results. Finally, as stated earlier, we also acknowledge the need for the empirical validation of the proposed framework and propositions. Our current work includes such empirical research including expert interviews and case studies.

References

Agarwal, P. (2020). Gender bias in STEM: Women in Tech still facing discrimination. *Forbes*. Retrieved from https://www.forbes.com/sites/pragyaagarwaleurope/2020/03/04/gender-bias-in-stem-women-in-tech-report-facing-discrimination/?sh=72c9e78670fb

Ahn, Y., Lin, Y.R. (2020). Fair sight: Visual analytics for fairness in decision making. *IEEE Transactions on Visualization and Computer Graphics*, 26 (1). doi: https://doi.org/10.1109/TVCG.2019.2934262

Altman, M., Wood, A., Vayena, E. (2018). A harm-reduction framework for algorithmic fairness. *IEEE Security & Privacy*, 16(3), 34-45.

Arrieta, A. B., Diaz – Rodriguez, N., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil- Lopez, S., Molina, D., Benjamins, R., Chatila, R., Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges towards responsible AI. *Information Fusion*, 58, 82-115.

Bandara, W., Miskon, S., Fielt, E. (2011). A systematic, tool-supported method for conducting literature reviews in information systems. In *Proceedings of the 19th European Conference on Information Systems,* Helsinki, Finland. https://aisel.aisnet.org/ecis2011/221

Baskerville, R. L., Myers, M. D., & Yoo, Y. (2020). Digital first: The ontological reversal and new challenges for IS. *MIS Quarterly*, 44(2), 509 -523. doi: 10.25300/MISQ/2020/14418

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., Zhang, Y. (2018). AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *Journal of Research and Development*, 1 (1), 99. doi: https://arxiv.org/pdf/1810.01943.pdf

Berente, N., Seidel, S., Safadi, H. (2019). Data-Driven Computationally-Intensive Theory Development. *Information Systems Research*, 30(1), 50-64.

Berente, N., Gu, B., Recker, J., Santhanam, R. (2019). Managing AI: Special issue, *MIS Quarterly*, 45(3), 1433-1450.

Benjamin, R. (2019). Assessing risk, automating racism. A health care algorithm reflects underlying racial bias in society. *Social Science.* doi: 10.1126/science.aaz3873

Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A. (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods and Research*. 50 (1). https://doi.org/10.1177/0049124118782533

Bolukbasi, T., Chang, K.W., Zou, J., Saligrama, V., Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Proceedings of the 30th International Conference on Neural Information Processing Systems* (pp. 4356–4364). https://dl.acm.org/doi/10.5555/3157382.3157584

Borges, A. F. S., Laurindo, F.J.B., Spinola, M. M., Goncalves, R. F., Mattos, C.A. (2021). The strategic use of artificial intelligence in the digital era: systematic literature review and future research directions. *International Journal of Information Management*, 57. doi: https://doi.org/10.1016/j.ijinfomgt.2020.102225

Canetti, R., Cohen, A., Dikkala, N., Ramnarayan, G., Scheffler, S., Smith, A. (2019). From soft classifiers to hard decisions: How far can we be? *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 309-318. doi: https://doi.org/10.1145/3287560.3287561

Cao, J., Basoglu, K., Sheng. H., Lowry, P. (2015). A systematic review of social network research in Information Systems: Building a foundation for exciting Future research. *Communications of the Association for Information Systems*, 36(37), 727-758.

Caplan, R., Donovan, J., Hanson, L., Matthews, J. (2018). Algorithmic Accountability: A Primer: Prepared for the Congressional Progressive Caucus. *Data & Society*, Washington, DC, USA. doi: https://datasociety.net/library/algorithmic-accountability-a-primer/

Chen, I. Y., Szolovits, P., Ghassemi, M. (2019). Can AI help reduce disparities in general medical and mental health care? *AMA Journal of Ethics*, 22 (2), 167- 179.

Cirillo, D., Catuara – Solarz, S., Morey, C., Guney, E., Subirats, L., Mellino, S., Gigante, A., Valencia, A., Rementeria, M. J., Chadha, A. S., Nikolaos, M. (2020). Sex and gender differences and biases in artificial intelligence for biomedicine and healthcare. *Digital Medicine*, 8 (3). doi: https://www.nature.com/articles/s41746-020-0288-5

Clifton, J., Glasmeier, A., Gray, M. (2020). When machines think for us: the consequences for work and place. *Cambridge Journal of Regions, Economy, and Society*,13(1), 3-23.

Costa, P., Ribas, L. (2019). AI becomes her: Discussing gender and artificial intelligence. *A Journal of Speculative Research*,17 (2), 171-193.

Conboy, K., Crowston, K., Lundstrom, J.E., Jarvenpaa, S., Ram, S., Mikalef, P. (2022). Artificial intelligence in Information systems: State of the art and research roadmap. *Communications of the Association for Information Systems,* 50, 420- 438.

Collins, C., Dennehy, D., Conboy, K., Mikalef, P. (2021). Artificial intelligence in information system research: A systematic literature review and research agenda. *International Journal of Information Management*, 60. doi: https://doi.org/10.1016/j.ijinfomgt.2021.102383

Crawford, K. (2016). A.I.'s White Guy Problem. (Sunday Review Desk) (OPINION). *The New York Times*. doi:https://www.nytimes.com/2016/06/26/opinion/sunday/artificial-intelligences-white-guy-problem.html

Dastin, J. (2018). Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters.* doi: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G

Daugherty, P., Wilson, H., Chowdhury, R. (2018). Using artificial intelligence to promote diversity. *MIT Sloan Management Review*. doi:https://sloanreview.mit.edu/article/using-artificial-intelligence-to-promote-diversity/

Dawson, D., Schleiger, E., Horton, J., McLaughlin, J., Robinson, C., Quezada, G., Scowcroft, J., Hajkowicz, S. (2019). Artificial Intelligence: Australia's Ethics Framework. *Data61 CSIRO*, Australia. doi: https://www.csiro.au/en/research/technology-space/ai/ai-ethics-framework

Dwivedi, Y.K., Hughes, L, …, Williams, M. D. (2019). Artificial Intelligence: Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management,* 57(7). doi:10.1016/j.ijinfomgt.2019.08.002

Eubanks, V. (2018). Automating inequalities: How high-tech tools profile, police, and punish the poor. *Law Technology and Humans.* doi: https://dl.acm.org/doi/10.5555/3208509

European Union (2021). Europe Fit for the Digital Age: Commission Proposes New Rules and Actions for Excellence and Trust in Artificial Intelligence. Available at https://ec.europa.eu/commission/presscorner/detail/en/ip_21_1682

Feuerriegel, S., Dolata, M., Schwabe, G. (2020). Fair AI: Challenges and opportunities. *Business and Information Systems Engineering*, 62(4), 379-384.

Feast, J., (2019). 4 ways to address gender bias in AI, *Harvard Business Review*. doi: https://hbr.org/2019/11/4-ways-to-address-gender-bias-in-ai

Galleno, A., Krentz, M., Tsusaka, M., Yousif, N. (2019). How AI could help or hinder women in the workforce. *Boston Consulting Group.* doi:https://www.bcg.com/publications/2019/artificial-intelligence-ai-help-hinder-women-workforce.

Grari. V., Ruf. B., Lamprier. S., Detyniecki. M. (2020). Achieving fairness with decision tress: An adversarial approach. *Data Science and Engineering*, 5(2). 99- 110.

Hardt, M., Price, E., Srebro, N. (2016). Equality of opportunity in supervised learning. In *Proceeding of the 30th International Conference on Neural Information Processing Systems.* doi: https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper

Hayes, P., Poel, I.V.D., Steen, M. (2020). Algorithms and values in justice and security. *AI & Society*, 35 (3), 533- 555.

Hoffmann, A.L. (2019). Where fairness fails: data, algorithms, and the limits of anti-discrimination discourse. *Information, Communication & Society*, 22(7*),* 900-915.

ICO (2018). What is automated individual decision-making and profiling. *UK Information Commissioner's Office*, 1-23. doi: https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-dataprotection-regulation-gdpr/automated-decision-making-and-profiling/what-is-automated-individual-decisionmaking-and-profiling/#id2

Ibrahim, S.A., Charlson, M.E., Neill, D.B. (2020). Big data analytics and the structure for equity in healthcare: The promise and perils. *Health Equity*, 4 (1), 99- 101.

Johnson, K.N. (2019). Automating the risk of bias. *George Washington Law Review*, 87(6). Doi: https://www.gwlr.org/wp-content/uploads/2020/01/87-Geo.-Wash.-L.-Rev.-1214.pdf

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

Kaplan, A., Haenlein, M. (2019). Siri, siri, in my hand: Who's the fairest in the land? On the interpretations, illustrations, and implication of artificial intelligence. *Business Horizons*, 62, 15-35.

Kitchenham, B., Budgen, D., Brereton, O.P. (2011). Using mapping studies as the basis for further research – A participant -observer case study. *Information and Software Technology,* 53 (6), 638- 651.

Kordzadeh, N., Ghasemaghaei, M. (2021). Algorithmic bias: review synthesis, and future research directions. *European Journal of Information Systems,* 31 (3), 388- 409.

Kyriazanos, D. M., Thanos, K.G., Thomopoulos, S.C.A. (2019). Automated decisions making in airports checkpoints: Bias detection toward smarter security and fairness. *IEEE Security & Applications*, 17 (2), 8-16.

Lambrecht, A., Tucker, C. (2019). Algorithmic bias? An empirical study of apparent gender-biased discrimination in the display of STEM career ads. *Management Science,* 65(7), 2947-3448.

Leavy, S. (2018). Gender bias in artificial intelligence: The need for diversity and gender theory in Machine learning. *2018 IEEE/ACM First international workshop on gender equality in software engineering,* Gothenburg, Sweden.
doi: https://ieeexplore.ieee.org/document/8452744

Lee, N. T. (2018). Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication, and Ethics in Society*, 16 (3), 252-260.

Marabelli, M., Newell, S., Handunge, V. (2021). The lifecycle of algorithmic decision- making systems: Organizational choices and ethical challenges. *Journal of Strategic Information Systems*, 30(3). doi: https://doi.org/10.1016/j.jsis.2021.101683

Markus (2017). Datification, organizational strategy, and IS research: What's the score? *The Journal of Strategic Information Systems*, 26(3), 233-241. doi:https://doi.org/10.1016/j.jsis.2017.08.003.

Martinez, C. F., Fernandez, A. (2020). AI and recruiting software: Ethical and legal implications. *Journal of Behavioural Robotics*, 11(1). doi: https://doi.org/10.1515/pjbr-2020-0030

Martin, K. (2019). Ethical implications and accountability of algorithms. *Journal of Business Ethics*, 160, 835- 850.

Marjanovic, O., Cecez-Kecmanovic, D., Vidgen, R. (2021). Algorithmic pollution: Making the invisible visible. *Journal of Information Technology*, 36(3), 391-408.

Masiero, S., Aaltonen, A. (2020). Gender bias in IS research: A literature review. In *Proceedings of the 41st International Conference on Information Systems*, Hyderabad, India. doi: http://dx.doi.org/10.2139/ssrn.3751440

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A. (2019). A survey on bias and fairness in machine learning. ACM Computing Surveys 54 (6). https://doi.org/10.1145/3457607

Mikolov, T, Chen, Corrado, G. S., Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In *Proceedings of the International Conference on Learning Representations*. Retrieved from https://storage.googleapis.com/pub-tools-public-publication-data/pdf/41224.pdf

Miron, M., Tolan, S., Gomez, E., Castillo, C. (2020). Evaluating causes of algorithmic bias in juvenile criminal recidivism. *Artificial Law and Intelligence*, 29(2), 111-147.

Nadeem, A., Abedin, B., Marjanovic, O. (2020). Gender bias in AI: A review of contributing factors and mitigating strategies. In *Proceedings of the Australasian Conference on Information Systems*, Wellington, New Zealand. https://aisel.aisnet.org/acis2020/27

Nadeem, A., Marjanovic, O., Abedin, B. (2021). Gender bias in AI: Implications for managerial practices. *13E 2021. Responsible AI and analytics for an ethical and inclusive digitized society.* doi: https://link.springer.com/chapter/10.1007/978-3-030-85447-8_23

Niehaus, F., Wiesche, M. (2021). A Socio-Technical perspective on organizational interaction with AI: A literature review. In *Proceedings of European Conference on Information Systems 2021*. doi: https://aisel.aisnet.org/ecis2021_rp/156

Ntoutsi, E., Fafalios, P., …, Staab, S., (2019). Bias in data-driven artificial intelligence systems – An introductory survey. *Data Mining and Knowledge Discovery*, 10(3). doi: https://doi.org/10.1002/widm.1356

Noriega., M. (2020). The application of artificial intelligence in police interrogations: An analysis addressing the proposed effect AI has on racial and gender bias, cooperation, and false confessions. *Futures*, 117. doi:10.1016/j.futures.2019.102510

Pare`, G., Trudel, M-C., Jaana, M., Kitsiou, S. (2015). Synthesizing information systems knowledge: A typology of literature reviews. *Information & Management*, 52(2) 183-199.

Parikh, R.B., Teeple, S., Navathe, A.M. (2019). Addressing bias in artificial intelligence in health care. *JAMA*. doi: https://pubmed.ncbi.nlm.nih.gov/31755905/

Parsheera, S. (2018). A gendered perspective on Artificial Intelligence. In *Proceeding of ITU Kaleidoscope – Machine learning for a 5G Future.* doi: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3374955

Petersen, K., Vakkalanka, S., Kuzniarz, J. (2015). Guidelines for conducting systematic mapping studies in software engineering: An update. *Information and Software Technology,* 64, 1-18.

Paulus, J. K., Kent, D. M. (2020). Predictably unequal: Understanding and addressing concerns that algorithmic clinical prediction may increase health disparities. *Digital Medicine*, 99 (3). https://doi.org/10.1038/s41746-020-0304-9

Piano, S.L. (2020). Ethical principles in machine learning and artificial intelligence: cases from the field and possible ways forward. *Humanities and Social Sciences Communications*. doi: https://www.nature.com/articles/s41599-020-0501-9

Prates, M., Avelar, P., Lamb, L.C. (2018). Assessing gender bias in machine translation – A case study with google translate. *Neural Computing and Applications*. 32, 6363 – 6381.

Qureshi, B., Kamiran, F., Karim, A., Ruggieri, S., Pedreschi, D. (2020). Causal inference for social discrimination reasoning. *Journal of Intelligent Information Systems*, 54, 425-437.

Robert, L.P., Pierce, C., Marquis, L., Kim, S., Alahmad, R. (2020). Designing fair AI for managing employees in organizations: a review, critique, and design agenda. *Human - Computer Interaction.* 35(5-6). doi: https://doi.org/10.1080/07370024.2020.1735391

Robnett, R. D. (2015). Gender bias in STEM fields: Variation in prevalence and links to STEM self-concept. *Psychology of Women Quarterly, 40(1)*. https://doi.org/10.1177/0361684315596162

Sarker, S., Chatterjee, S., Xiao, X., Elbanna, A. (2019). The socio technical axis of cohesion for the IS discipline: Its historical legacy and its continued relevance. *MIS Quarterly*, 43(3), 695-719.

Sun, T., Gaut, A., Tang, S., Huang, Y., Elsherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K. W., Wang, W. Y. (2019). Mitigating gender bias in Natural Language Processing: A literature review. In *Proceeding of* 57th *Annual Meeting of Association for Computational Linguistics*, Florence, Italy. doi: 10.18653/v1/P19-1159

Soleimani, M., Intezari, A., Pauleen, D.J. (2021). Mitigating cognitive biases in developing AI-assisted recruitment systems: A knowledge-sharing approach. *International Journal of Knowledge Management,* 18(1). Doi: 10.4018/IJKM.290022

Schonberger, D. (2019). Artificial intelligence in healthcare: a critical analysis of the legal and ethical implications. *International Journal of Law and Information Technology*, 27 (2), 171-203.

Teodorescu, M., Morse, L., Awwad, Y., Kane. G. (2021). Failures of Fairness in Automation Require a Deeper Understanding of Human-ML Augmentation. *MIS Quarterly*, 45(3), 1483-1500.

Thelwall, M. (2017). Gender bias in machine learning for sentiment analysis. *Online Information Review*, 42(3), 343- 354.

Trewin, S., Basson, S., Muller, M., Branham, S., Treviranus, J., Gruen, D., Hebert, D., Lyckowski, N., Manser, E. (2019). Considerations for AI fairness for people with disabilities. *AI Matters,* 5(3). doi: 10.1145/3362077.3362086

United Nations Educational, Scientific and cultural organization (UNESCO) (2020). Artificial intelligence, and gender equality: key finding of UNESCO's global dialogue. Available at https://unesdoc.unesco.org/ark:/48223/pf0000374174

Veale, M., Binns, R. (2017). Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data and Society*, 1-17. doi: https://doi.org/10.1177/2053951717743530

Wang, L. (2020). The three harms of gendered technology. *Australasian Journal of Information Systems*, 24. doi: https://doi.org/10.3127/ajis.v24i0.2799

West, S.M., Whittaker, M., Crawford, K. (2019). Discriminating systems: Gender, race and power in AI. *AI Now Institute.* Retrieved from https://ainowinstitute.org/discriminatingsystems.html

Webster, J., Watson, R.T. (2002). Analysing the past to prepare for the future: Writing a literature review, *MIS Quarterly*, 26(2), 3-23.

Wolfswinkel, J., Furtmueller, E., Wilderom, C. (2013). Using grounded theory as a method for rigorous reviewing literature. *European Journal of Information Systems*, 22 (1), 45-55.

Wu, W., Huang, T., Gong, K. (2019). Ethical principles and governance technology development of AI in China. *Engineering*, 6(3), 302-309.

Zhong, Z. (2018). A tutorial on fairness in machine learning, *Towards Data Science*. https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040

Appendix

Table 1. Articles selected for the literature review

Journal Name	Title of the article	Author	Theory used	Type of article	Focus
Data Science and Engineering	Achieving fairness with decision trees	Grari et al., 2020	No theory used	Design science	Decision-making (decision trees)
Journal of Behavioural Robotics	AI & recruiting software: Ethical and legal implications	Martinez & Fernandez, 2020	No theory used	Conceptual and essays.	Human resource
Journal of Speculative Research	AI becomes her: Discussing gender and AI	Costa & Ribas, 2019	Gender theory	Conceptual and essays	Gendered technology

Journal of Research and Development	AI fairness 360	Bellamy, 2018	No theory used	Design science	Fairness toolkit
AI & Society	Algorithms and values injustice & security	Hayes, Poel, Steen, 2020	No theory used	Conceptual and essays	Societal
International Journal of Law and Information Technology	AI in healthcare: a critical analysis	Schonberger, 2019	No theory used	Literature review	Healthcare industry sector
Neural Computing and Applications	Assessing gender bias in Machine translation	Prates, Avelar, Lamb, 2019	No theory used	Empirical (Case study)	Machine learning
IEEE Security and Privacy	Automated decision making in airport checkpoints:	Kyriazanos et al., 2019	No theory used	Conceptual and essays.	Airport checkpoints
George Washington Law review	Automating the risk of bias	Johnson, 2019	Risk management theory, Traditional agency theory, Theory of commitment bias	Conceptual and essays	AI software tools
Management Science	Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM Career	Lambrecht & Tucker, 2019	No theory used	Design science	STEM career
Data mining and knowledge discovery	Bias in data-driven AI systems	Ntoutsi et al., 2019	No theory used	Empirical (Survey)	Data
Health Equity	Big data analytics and the struggle for equity in healthcare: The promise & perils.	Ibrahim, Charlson, Neill, 2020	No theory used	Conceptual and essays.	Healthcare industry sector
AMA Journal of Ethics	Can AI help reduce disparities in General Medical & Medical healthcare?	Chen, Szolovits, Ghassemi, 2019	No theory used	Empirical (Case study)	Healthcare industry sector
Journal of Intelligent Information Systems	Causal inference for social discrimination reasoning.	Qureshi et al., 2020	No theory used	Design science	Testing of algorithms
Human-Computer Interaction	Designing fair AI for managing employees in organizations: A review, critique & design agenda	Robert et al., 2020	Organizational justice theory & Adam's equity theory	Literature review	Human resource

Journal of Information, Communication, and Ethics in Society	Detecting racial bias in algorithm and machine learning	Lee, 2018	No theory used	Conceptual and essays.	Machine learning
Journal of Business Ethics	Ethical implications and accountability of algorithms	Martin, 2019	Theory of algorithmic accountability, Decision-making theory	Conceptual and essays.	Ethics & accountability of algorithms
Engineering	Ethical principles and governance technology development of AI in China	Wu, Huang, Gong, 2019	No theory used	Conceptual and essays.	Country (national use of AI)
Humanities and Social Sciences Communications	Ethical principles in ML & AI: Cases from the field and possible ways forward	Piano, 2020	No theory used	Literature review	Autonomous vehicles
Artificial intelligence and Law	Evaluating causes of algorithmic bias in Juvenile criminal recidivism	Miron et al., 2020	No theory used	Designs science	Criminal justice
Information Fusion	Explainable AI(XAI): Concepts, taxonomies, opportunities & challenges towards responsible AI	Arrieta et al., 2020	Game theory, Theory-guided data science	Conceptual and essays	Responsible AI
Business and Information systems Engineering	Fair AI - Challenges and opportunities	Feuerriegel, Dolata, Schwabe, 2020	No theory used	Conceptual and essays.	Fairness in AI
Big Data and Society	Fair machine learning in the real world: mitigating discrimination without collecting sensitive data	Veale & Binns, 2017	No theory used	Conceptual and essays.	Fairness in AI
Sociological Methods and Research	Fairness in criminal justice risk assessments: The state of the art.	Berk et al., 2018	No theory used	Literature review	Criminal justice
Online Information Review	Gender bias in ML for sentiment analysis	Thelwall 2017	No theory used.	Design science	ML
Digital Medicine	Predictably unequal: Understanding and addressing concerns that	Paulus & Kent, 2020	No theory used	Design science	Healthcare

	algorithmic clinical predictions increase health disparities				
Digital Medicine	Sex & gender differences and biases in AI for biomedicine & healthcare	Cirillo et al., 2020	No theory used	Empirical (Survey)	Healthcare
Futures	The application of AI in police interrogations: An analysis addressing the proposed effect AI has on racial and gender bias	Noriega, 2020	Uncanny theory	Conceptual and essays	Criminal justice
Australasian Journal of Information Systems	The Three Harms of Gendered Technology	Wang, 2020	No theory used	Conceptual and essays	Decision making
IEEE Transactions on Visualization and Computer Graphics	Fair Sight: Visual Analytics for Fairness in Decision Making	Ahn & Lin, 2020	No theory used	Empirical (Case study)	Visual analytics
Cambridge Journal of Regions, Economy, and Society	When machines think for us: the consequences for work and place	Clifton, Glasmeier, Gray, 2020	No theory used	Conceptual and essays	Workplaces and society

Table 2. Characteristics of gender bias in AI

Grouping of concepts/ themes	Grouping of characteristics/concepts	Characteristics of gender bias in AI	Source	Description
Societal	Prejudices in society	Societal gender prejudices	Martinez & Fernandez 2020; Johnson 2019; Ntoutsi et al., 2019; Thelwall, 2017; Noriega, 2020; Wang, 2020; Clifton, Glasmeier, Gray 2020; Prates, Avelar, Lamb 2019; Lee, 2018; Cirillo et al., 2020	Pre-existing societal inequalities, such as internalized misogyny.
		Discrimination	Grari et al., 2020; Lambrecht & Tucker, 2019	
	Biased behaviours followed by the majority	Pre-existing norms of society are followed by the majority	Ntoutsi et al., 2019	

Institutional	Biases across various disciplines – socio-demographic & technological biases	Association of femineity with certain soft skills (non-technical)	Costa & Ribas, 2019	Socio-economic factors contribute to discrimination.
		Socio-economic factors requiring interdisciplinary collaboration	Ibrahim, Charlson, Neill, 2020; Ntoutsi et al., 2019, Martin, 2019; Veale & Binns 2017	
Design & Implementation	Lack of gender diversity in data and technology	Lack of gender diversity in AI development and training datasets	Martinez & Fernandez, 2020; Johnson, 2019; Wang, 2020; Clifton, Glasmeier, Gray 2020	Existing issues in societal bias sneak into the design and implementation of technology.
	AI amplifies the bias in society by producing biased outcomes	AI amplifying social prejudices	Grari et al., 2020; Hayes, Poel, Steen 2020; Johnson, 2019	Algorithms amplify those phenomena that are easily quantifiable.
		Nascent technology creates a risk	Johnson 2019	
	Prejudices influencing technology through biased data	Pre-existing patterns of exclusions and disparities discovered by data mining	Johnson, 2019; Clifton, Glasmeier, Gray 2020	Biased data seeps into the AI algorithms resulting in amplifying discrimination and inequalities in societies.
		Disparities in society embedded in data	Ibrahim, Charlson, Neill, 2020; Veale, Binns 2017	

Table 3. Contributing factors of gender bias in AI

Grouping of concepts/ themes	**Grouping of factors/ concepts**	**Contributing factors of gender bias in AI**	**Source**	**Description**
Biased training datasets	Misrepresentation of subjects in training datasets	Improper data gathering practices	Grari et al., 2020; Martinez, Fernandez, 2020; Hayes, Poel, Steen, 2020; Kyriazanos et al., 2019; Johnson 2019; Lambrecht, Tucker, 2019; Ntoutsi et al., 2019; Chen, Szolovits, Ghassemi, 2019; Ibrahim, Charlson, Neill, 2020; Qureshi et al., 2020; Lee, 2018; Martin, 2019; Miron et al., 2020; Arrieta et al., 2020; Feuerriegel, Dolata, Schwabe, 2020; Thelwall, 2017; Paulus, Kent, 2020; Cirillo et al., 2020; Noriega, 2020; Ahn, Lin, 2020;	Over and under-representation of certain groups in data sets can result to perpetuate discrimination. Datasets may be underrepresented of the public demographics.

			Bellamy, 2018; Feuerriegel, Dolata, Schwabe, 2020	
		Under or over-representation of subjects in datasets	Martinez, Fernandez, 2020; Hayes, Poel, Steen, 2020; Johnson, 2019; Ntoutsi et al., 2019; Robert et al., 2020; Martin, 2019; Miron et al., 2020; Veale, Binns, 2017; Cirillo et al., 2020; Clifton, Glasmeier, Gray, 2020	
		Unavailability of useful data	Hayes, Poel, Steen, 2020; Veale, Binns, 2017; Noriega, 2020	
	Unfair training datasets	Language discrimination for gender in data	Prates, Avelar, Lamb, 2019; Ntoutsi et al., 2019; Chen, Szolovits, Ghassemi, 2019; Qureshi et al., 2020; Lee, 2018; Thelwall, 2017; Cirillo et al., 2020	Patterns in the data are designed to discriminate.
		Unfairness in data	Veale, Binns, 2017.	
	Programmers/data miners' conscious or unconscious bias	Data miners unintentionally parse the bias while discovering patterns of inequalities in data	Hayes, Poel, Steen, 2020; Johnson, 2019; Ntoutsi et al., 2019, Lee, 2018; Martin, 2019.	Programmers unintentionally incorporate bias during the input, training, and programming stage.
		Programmers' conscious or unconscious bias	Hayes, Poel, Steen, 2020; Johnson, 2019; Lambrecht, Tucker, 2019; Piano, 2020; Noriega, 2020, Wang, 2020; Clifton, Glasmeier, Gray, 2020	
	Proxy variables of sensitive features	Variables acting as a proxy - sensitive features and their casual influences in data	Martinez, Fernandez, 2020; Ntoutsi et al., 2019; Ibrahim, Charlson, Neill, 2020; Robert et al., 2020; Lee, 2018; Martin, 2019; Piano, 2020; Arrieta et al., 2020, Feuerriegel, Dolata, Schwabe, 2020; Noriega, 2020; Ahn, Lin, 2020; Bellamy, 2018; Robert et al., 2020.	Proxy data are being used for features that are hard to quantify or to be collected.
		Co-relational analysis of observational data	Qureshi et al., 2020	
	Historical human bias in data	Historical bias goes to biased datasets	Hayes, Poel, Steen, 2020; Kyriazanos et al., 2019; Johnson, 2019; Lambrecht, Tucker, 2019; Ibrahim, Charlson, Neill, 2020; Martin, 2019; Veale, Binns, 2017; Cirillo et al., 2020; Ahn, Lin, 2020.	The majority follow the norms established by the society in the past.

Gender stereotyping	Prejudices in society	Gender stereotyping in society	Prates, Avelar, Lamb, 2019; Lee, 2018; Miron et al., 2020; Cirillo et al., 2020; Noriega 2020; Wang 2020	Prevailing gender stereotyping in society.
		Societal prejudices	Martinez, Fernandez, 2020; Ntoutsi et al., 2019; Miron et al., 2020; Thelwall, 2017; Cirillo et al., 2020; Noriega, 2020.	
	Culture fostering masculinity	Bro culture fostering exclusivity and masculinity	Johnson, 2019	Gender imbalance and masculinity are pervasive in the IT industry.
	Socio-economic factors imputing discrimination	economic factors	Martinez, Fernandez, 2020; Hayes, Poel, Steen, 2020; Lambrecht, Tucker, 2019; Cirillo et al., 2020; Wang, 2020.	Socio-economic factors based on social standing e.g. neighbourhood, zip code, location results in incorrect assumptions of an individual.
		Individual socio-status based on zip code	Lee, 2018; Veale & Binns, 2017	
	Decisions are taken on biased norms followed by the majority	Decisions are made on pre-existing norms of society followed by the majority	Ntoutsi et al., 2019	Certain upstream social norms are followed blindly because of their easy acceptance in society.
AI amplifies existing bias	Failing to ensure humans in the loop in AI decisions	Reducing the role of human agents or failing to ensure "human in the loop"	Johnson, 2019	Data mining systems reproduce the historic biases embedded in the data if there is a missing human role in the final decision making
		Algorithms optimizing cost-effectiveness in a discriminatory way	Lambrecht & Tucker, 2019; Ntoutsi et al., 2019; Chen, Szolovit, Ghassemi, 2019; Ibrahim, Charlson, Neill, 2020; Martin, 2019.	
	AI amplifying the bias in society	The algorithm perpetuates/ amplifies discrimination and biases	Hayes, Poel, Steen, 2020; Johnson, 2019, Ntoutsi et al., 2019; Lee, 2018; Cirillo et al., 2020; Clifton, Glasmeier, Gray, 2020; Clifton, Glasmeier, Gray 2020	Algorithms inherent rules from previously discriminatory decisions.
	Lack of transparency in algorithms	Opacity in algorithms	Hayes, Poel, Steen, 2020; Miron et al., 2020; Ahn, Lin, 2020; Clifton, Glasmeier, Gray, 2020.	Algorithms are opaque, complex, unpredictable, and partially autonomous.

	Creator's inherent bias in AI algorithms	AI reflecting creator's bias	Hayes, Poel, Steen, 2020; Ntoutsi et al., 2019, Miron et al., 2020; Wang, 2020.	The lack of multidisciplinary aspects in AI creators results in unfair outcomes.
Lack of gender diversity in AI development teams	Lack of Gender disparity in developers and the technology sector	Gender disparity in AI development team	Clifton, Glasmeier, Gray, 2020; Martinez, Fernandez, 2020; Johnson, 2019; Wang, 2020	There is a lack of diversity in AI developers' teams and in the technology sector because of which there is a lack of diversity of thought in the preparation of the data as unconscious bias seeps in the training data in the data preparation stage.
		Lack of diversity in the STEM sector in senior management and employees.	Johnson, 2019; Lee, 2018; Wang, 2020.	
		The technology industry is remarkably male-dominated and exceptionally homogeneous.	Johnson, 2019; Lee, 2018; Wang, 2020	
Lack of AI regulations	Limited regulation on data collection, selection, and modification	Lack of legal provision dealing with the way data is collected, selected, and modified	Ntoutsi et al., 2019; Wang, 2020	Data protection laws, and general provisions concerning data quality are deficient.
	Limited regulations addressing gender bias	Limited development in regulation towards addressing gender balance.	Johnson, 2019.	
Contextual and other factors	Agents behind the data collection lack awareness of certain sensitive features.	Agents such as executives, software engineers, data scientists, developers, and policymakers commissioning and authorizing the algorithm are not aware of sensitive features	Hayes, Poel, Steen, 2020; Wang, 2020.	AI development teams are not aware of the importance of distinguishing between certain categories.
		Developers may fail or ignore to specify the limits of datasets	Johnson 2019; Ntoutsi et al., 2019.	
		The third party that has collected the underlying dataset may aggregate data	Johnson, 2019.	

	Algorithms not tested for a specific context/application/ sector	Dumb-start programs that are not designed or tested for a specific context	Qureshi et al., 2020.	Failing to test an algorithm with regards to the context in which it would be used.

Table 4. Approaches for addressing gender bias in AI

Grouping of conncepts/ themes	Grouping of factors/ concepts	Approaches for mitigating gender bias in AI	Source	Description
AI technology- related approaches	Collection and preparation of data sets	Bias-aware data collection	Ntoutsi et al., 2019; Bellamy, 2018	Preventing unfairness in training data by ensuring fair data collection, data preparation, and regularizing the training data to minimize the unfairness.
		Preparation of the fair data	Hayes, Poel, Steen, 2020; Qureshi et al., 2020; Arrieta et al., 2020; Noriega, 2020; Grari et al., 2020; Kyriazanos et al., 2019; Ntoutsi et al., 2019; Miron et al., 2020; Arrieta et al., 2020; Veale, Binns, 2017; Berk et al., 2018; Ahn, Lin, 2020; Bellamy, 2018.	
		Removing proxies of protected attributes from the datasets.	Grari et al., 2020; Hayes, Poel, Steen, 2020; Miron et al., 2020; Arrieta et al., 2020; Feuerriegel, Dolata, Schwabe, 2020; Veale, Binns, 2017; Bellamy, 2018.	
	In-processing of algorithms	Integration of algorithm	Grari et al., 2020; Kyriazanos et al., 2019; Ntoutsi et al., 2019; Miron et al., 2020; Arrieta et al., 2020; Veale, Binns, 2017; Berk et al., 2018; Ahn, Lin, 2020; Bellamy, 2018.	Fair algorithmic integration and resource allocation to ensure strengthening of algorithmic design.
		Equal/unbiased resources of allocation in an algorithm	Grari et al., 2020; Lambrecht, Tucker, 2019; Lee, 2018; Miron et al., 2020; Arrieta et al., 2020; Veale, Binns, 2017; Ahn, Lin, 2020.	

		Designing fair classification of algorithms	Grari et al., 2020; Ntoutsi et al., 2019; Robert et al., 2020; Kyriazanos et al., 2019.	
		Strengthening of formal & statistical foundations of algorithms	Ntoutsi et al., 2019.	
	Implementation of algorithms	Interpreting & testing of the algorithms	Grari et al., 2020; Kyriazanos et al., 2019; Ntoutsi et al., 2019; Miron et al., 2020; Arrieta et al., 2020; Veale, Binns, 2017; Berk et al., 2018; Ahn, Lin, 2020; Bellamy, 2018.	Testing the algorithm for a specific application for enhanced accountability and bias detection.
		Ensuring algorithmic transparency, explain ability, and accountability	Johnson, 2019; Lambrecht, Tucker, 2019; Martin, 2019; Feuerriegel, Dolata, Schwabe, 2020; Thelwall, 2017; Clifton, Glasmeier, Gray, 2020; Arrieta et al., 2020; Ntoutsi et al., 2019; Cirillo et al., 2020.	
Fair AI management approaches	Better fairness governance policies	Internal governance policies	Johnson, 2019.	Enhanced AI corporate governance for gender bias mitigation
		Internal structures and process-oriented corporate governance	Johnson, 2019; Martin, 2019	
	Continuous education/training on fairness and ethics for all stakeholders	Educational workshops and training on workplace fairness	Noriega, 2020.	Workshops/education that involves principles of ethics such as promoting ethical education for every stakeholder in AI research & development.
		Certified professional required	Martin, 2019.	
		Awareness of ethics and promoting responsible AI	Wu, Huang, Gong, 2019; Veale, Binns, 2017	

		Awareness of unintended bias in scientific community and technology industry	Cirillo et al., 2020	
	Collaborative organizational learning on fairness & demographic characteristics	Business models and policy should be designed concerning fair AI	Feuerriegel, Dolata, Schwabe, 2020	Design of business models and policies to consider AI principles.
	Interdisciplinary approach & understanding of AI ethical principles	Interdisciplinary disciplines to work collaboratively to address ethical challenges	Wu, Huang, Gong, 2019; Ibrahim, Charlson, Neill, 2020	Employment of a more diverse IT workforce to be included in the design and implementation of algorithms.
	Workplace diversity in managerial roles	Gender diversity at managerial levels	Lee, 2018	An increase in gender inclusion in the development of AI technologies will introduce diverse perspectives and diversity of thought in the AI development which is essential for breaking down the bias.
		Diversity in the development of AI systems	Costa, Ribas, 2019; Johnson 2019; Ntoutsi et al., 2019; Arrieta et al., 2020; Clifton, Glasmeier, Gray 2020	
		Gender diversity in the high-tech industry and STEM career	Lee, 2018; Johnson, 2019; Wang, 2020	
	Designing strategies for incorporating algorithmic transparency and accountability	Big data review board required	Martin, 2019	AI audits are to be conducted periodically to ensure AI compliance.
		Incorporate regular audits of the data	Martinez, Fernandez, 2020; Johnson, 2019; Ibrahim, Charlson, Neill, 2020; Robert et al., 2020; Piano, 2020; Veale, Binns, 2017; Noriega, 2020	
		Designing strategies for fairness and ensuring accountability	Hayes, Poel, Steen, 2020	
	Ensuring Human in the loop	Integrating human & AI decision making	Miron et al., 2020	Design strategies like providing more autonomy to the users in decision-making would bring

				fairness to AI decisions.
AI governance and regulatory approaches	AI governance to incorporate key ethical standards	Ethical regulation for new technology development	Clifton, Glasmeier, Gray, 2020; Wang, 2020	Ethical standards by government and regulatory organizations to ensure fairer data collection and models
		Public policy regarding fair AI	Clifton, Glasmeier, Gray, 2020	
		Ethical technology development	Cirillo et al., 2020	
	Laws and policies to adhere to ethical AI principles	Quality assurance for AI safety.	W, Huang, Gong, 2019	Formal verification and testing to be carried out by users for ensuring AI safety and to adhere to AI ethical values.
		Algorithm test & trail	Lee, 2018; Wu, Huang, Gong, 2019	
		Practitioners and policymakers adhering to ethical AI principles	Lee, 2018	
		Regulated policies for implicit and unconscious bias	Lee, 2018	
	Ethical AI platforms	Comprehensive open AI platforms for all stakeholders	Wu, Huang, Gong, 2019	Developing open AI platforms for designing strategies or technical inputs to be incorporated to promote autonomy.
cietal & community focused approaches	Awareness of gender diversity	Logical considerations for increasing gender diversity	Hayes, Poel, Steen, 2020	Social intervention in order to enhance gender diversity in social and contextual aspects would address the gender bias in AI.
		Gender-neutral expression in communication	Prates, Avelar, Lamb, 2019	
		Bringing Cognitive diversity in workplace /institutions	Johnson, 2019	
	Gender diversity related to socio-economic aspects	Gender diversity related to socio-economic aspects	Cirillo et al., 2020	Compounding factors related to socio-economic aspects to

				establish fairness in AI algorithm "Ecosystem of trust" that ensures AI systems incorporate ethical standards
	Policy intervention	Policies to reach beyond the social prejudices	Clifton, Glasmeier, Gray, 2020	Ensuring public policy to bring trust for AI systems and societal wellbeing and fairness.
		Better AI policies for gender bias affectee	Clifton, Glasmeier, Gray, 2020	
		To comply with the rules that protect fundamental rights	Miron et al., 2020	

doi: https://doi.org/10.3127/ajis.v26i0.3835